RICE UNIVERSITY

Protein functional features extracted from primary sequences: a focus on disordered regions.

by

Natalia Pietrosemoli

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED THESIS COMMITTEE:

Jianpeng Ma, Chair
Professor, Department of Bioengineering

Laura Segatori,
Assistant Professor, Department of
Chemical and Biomolecular Engineering

Florencio Pazos,
Staff Scientist, Computational Systems
Biology Group. National Centre for
Biotechnology (CNB-CSIC)

Oleg Igoshin
Assistant Professor, Department of
Bioengineering

HOUSTON, TEXAS
DECEMBER 2012

ABSTRACT

Protein functional features extracted from primary sequences: a focus on disordered regions.

by

Natalia Pietrosemoli

In this thesis we implement an ensemble of sequence analysis strategies aimed at identifying functional and structural protein features. The first part of this work was dedicated to two case studies of specific proteins analyzed to provide candidate functional positions for experimental validation: the protein alpha-synuclein (αsyn) and the alanine racemases protein family. In the case of αsyn, the objective was to predict its aggregation prone regions. For the alanine racemase protein family, the scope was to predict sites responsible for substrate specificity. In these two studies, computational predictions allowed systematically exploring potentially functionally relevant protein sites in an efficient manner that may not be possible to implement with traditional experimental approaches. Our strategy provided a powerful forecasting tool for the selection of candidate sites to be later verified experimentally.

In the second part, we analyze the role of intrinsic disorder (ID) as a modulator of protein function in different organisms and cellular processes, which is largely unexplored. As key components of the diverse cellular pathways, disordered proteins are often involved in many diseases, including cancer and

neurodegenerative diseases. Thus, there is an impeding need to unveil the general principles underlying the role of ID in proteins. We provide a multi-scale analysis of the involvement of ID in protein function starting with a large-scale analysis at genomic level of the role of ID in *Arabidopsis*, zooming in into the specific processes of vesicular trafficking in Human and yeast, and finally focusing on specific proteins of diverse organisms.

The results of this thesis provide a better understanding of the functional roles mediated by ID in different organisms and biological processes, such as acting as flexible linkers connecting structured domains, mediating protein-protein interactions, and assisting the quick assembly of large macromolecular complexes. In addition, we present evidence of the use of ID as a mechanism to increase the complexity of protein and biological networks, and as a means to increase the adaptability of proteins in specific processes. Thus, our results contribute to elucidating the relationship between network and organismal complexity and ID, while they also provide evidence of the evolutionary advantages offered by ID.

Acknowledgements

I am lucky to say that I have received support and encouragement from a great number of people during these last years. The completion of my dissertation has not only been a very long journey, but quite a bumpy road too.

I would like to give special thanks to my Thesis Co-director, Dr. Florencio Pazos, head of the Computational Systems Biology Group at the National Centre for Biotechnology (CNB-CSIC), for giving me the opportunity of developing this dissertation under his guidance. He was full of invaluable suggestions and insightful comments, and it was truly a pleasure to work with him.

I would also like to thank my Thesis Co-director Dr. Laura Segatori, head of the Cell and Protein Engineering Laboratory at Rice University, not only for her excellent academic guidance, but also for her constant encouragement and great advice along all these years. Thanks for pushing me!

I also thank other members of my dissertation committee: Dr. Jianpeng Ma and Dr. Igoshin for their time and constructive feedback. I want to thank Dr. Rebecca Richards-Kortum, Chair of the Department of Bioengineering for her trust and support, and to the Department Coordinator, Gayle Schroeder for her constant help during my time abroad.

My most sincere gratitude goes to Prof. Alfonso Valencia, Director of the Structural Computational Biology and Biocomputing Programme at the Spanish

National Cancer Research Centre (CNIO), for giving me the privilege of joining his group as a visiting scholar and for his willingness to help in difficult times.

All my collaborators should also be thanked for the many interesting discussions and for making possible a good part of my research: Dr. Roberto Solano, head of the The jasmonate signalling pathway in Arabidopsis Group at the CNB, Juan Antonio Garcia, from the Department of Plant Molecular Genetics at the CNB, Dr. Peter Tompa, Chair of the Department of Structural Biology at the Vrije Universiteit Brussel, and Dr. Felipe Cava and Akbar Espaillat from the Department of Virology and Microbiology at the Severo Ochoa Molecular Biology Center (CBMSO-CSIC). I feel very lucky for having had the opportunity to work with people with such a broad spectrum of expertise.

I would like to thank the members of the Computational Systems Biology Group and the Bioinformatics Initiative at the CNB: Mónica, Toño, Dany, David O., Juan Carlos O., Juan Carlos S. and David SL. Closely working with such brilliant people has been a very inspiring intellectual experience. My thanks also go to all the members of my former Structural Computational Biology Group in the CNIO. I'll take with me great knowledge and invaluable memories of those times.

My love and gratefulness goes to all my family, especially to my wonderful parents and brother for all their love, patience and all kinds of support. Thanks to my granma, because she has always managed to transmit to me all her encouraging

words, tenaciousness and good vibes, wherever and whenever I need them the most. Granma, you truly rock.

To the rest of my big, big extended family of friends, past and present, you know you all contributed to this project in your own way, thank you! To my friend Chela, because.

To my big love Antonio, no words can express my gratitude for your unconditional support during these years. All I know is that if we could survive not one, but two of these “bumpy roads” together, we can do anything. We are a team!

Above all, I thank the Universe for offering me this wonderful opportunity to undertake this venture and enabling me to its completion.

Contents

List of Figures

List of Tables

Nomenclature

αsyn	Alpha synuclein
BrsV	Broad Spectrum Racemase in Vibrio
ID	Intrinsic Disorder
IDP	Intrinsically Disordered Proteins
IDR	Intrinsically Disordered Regions
MCA	Multiple Correspondence Analysis
MSA	Multiple Sequence Alignment
NN	Neural Networks
PCA	Principal Component Analysis
PLP	Pyridoxal 5”-Phospate
PPI	Protein-Protein Interaction
SDP	Specificity Determining Position
VM	Support Vector Machines

Introduction

One of the great challenges of the post-genomic era is to provide computer-based methods to interpret genomic data resulting from massive sequencing initiatives and from novel experimental techniques in molecular biology[1]. This quest aims at providing a better understanding of biological systems on organismal and cellular level. At the same time, there is a strong demand for immediate solutions, since deciphering the biological information encoded in genomic data will inevitably lead to important scientific findings. As the number of organisms successfully sequenced increases notably due to the availability of next-generation sequencing (NGS) technologies, so does the need for characterizing their encoded proteins. This thesis presents an ensemble of strategies for extracting functional and structural features from protein sequences, especially focusing on those features related to intrinsic disorder (ID).

In the first part of this thesis, we discuss two case studies of specific proteins in which we integrate computational tools with expert knowledge to analyze protein sequences and predict functional regions with the ultimate goal to guide experimental analyses. In the first study, the aim is to identify aggregation-prone regions of the alpha synuclein (αsyn) protein. The accumulation and aggregation of αsyn leads to the development of neurodegenerative diseases[2], and αsyn's aggregation is considered the hallmark Parkinson's disease[3,4]. The molecular

mechanisms that govern αsyn aggregation are not fully understood, and elucidating them can be fundamental for developing mechanisms to control aggregation. In this study, we propose a strategy to investigate the relationship between sequence-structure of αsyn and its aggregation propensity. We also propose some rationally designed protein variants with predicted effects on aggregation for experimental validation.

In the second study, the objective is to identify the molecular basis of substrate specificity of the alanine racemase protein family. Racemases are enzymes that convert L-amino acids into D-amino acids[5]. Alanine racemases are poorly characterized, and several alanine racemases have shown broader substrate specificity than what their current annotations report[6]. A better characterization of the enzymes and mechanisms involved in the synthesis and metabolism of noncanonical D-amino acids (NCDAAs) will also contribute to better understand NCDAAs production and their emerging roles in bacterial physiology. Thus, we propose to identify substrate specificity-determining positions in the alanine racemase protein family. Additionally, we propose novel racemase variants with specific substrate-binding affinities to be experimentally validated. In these two studies, computational predictions allow a systematic exploration of potentially functionally relevant protein sites in an efficient manner that may not be possible or difficult to implement with traditional experimental approaches. In fact, in both studies, protein variants have been experimentally tested and computational predictions confirmed.

In the second part of this thesis, we present a series of studies aimed at getting insights into the physiological function and functional mode of intrinsically disordered proteins and intrinsically disordered protein regions (IDPs/IDRs) at different scales. Over the past years there has been an increasing appreciation for the involvement of IDPs/IDRs in many signaling and regulatory cascades in protein interaction networks of eukaryotic cells[7,8,9]. In fact, the occurrence of IDPs/IDRs in key cellular signaling and regulatory processes is growing as genome sequences of multiple organisms become available and are inspected[10,11]. Moreover, being components with key cellular functions, IDPs are often linked to diverse pathologies including cancer, cardiovascular and neurodegenerative diseases[12,13,14,15,16]. Thus, there is growing interest in understanding IDP's abundance, molecular implications and function in the cell. Additionally, results from this study may contribute to explore the nascent use of IDPs as potential drug targets[8,7,17].

The evolution of the ID field has required the combined efforts of structural, molecular and systems biologists. However, it is largely due to computational predictions of ID if we now appreciate that IDPs are abundant and widespread and the cellular processes in which they are enriched. The ID field is developing at rapid pace, and even if many concepts such as ID's prevalence, functional roles and advantages have been accurately foreseen, they need to be systematically addressed. Ongoing efforts to study IDPs/IDRs biological functions include a variety of experimental techniques to characterize single proteins as well as computational

tools[18,19]. Computational predictions are crucial when analyzing whole organisms, systems that are too large or too heterogeneous for experimental characterization.

We present a multi-scale analysis of the implications of ID in protein function, starting with a large-scale analysis at genomic level of the role of ID in *Arabidopsis thaliana*, zooming in into the specific processes of vesicular trafficking in Human and yeast, and finally focusing on specific proteins of diverse organisms. We performed the first genome-wide analysis of ID in proteins from the model organism *Arabidopsis thaliana*, to identify disorder's functional roles in the underlying biological processes of this organism. We hypothesized that *A. thaliana* might heavily rely on disordered regions of proteins to respond to changes in environmental conditions and mediate the corresponding responses. Because plants are sessile organisms, they cannot escape from threatening conditions. As a result, plants depend on their phenotypic plasticity (i.e. the capacity to adapt their phenotype to changing conditions) to adapt and survive in rapidly changing environmental conditions. Implementing phenotypic plasticity requires the integration of external information with the basal genetic and developmental programs, which is achieved in plants through complex singling networks to which ID might be adding flexibility[20].

We then evaluate ID for proteins involved in vesicle trafficking. We assess the involvement of ID in proteins belonging to the three main vesicular trafficking routes in Human and yeast: the clathrin, coat protein complex I (COPI) and coat protein comple II (COPII) mediated routes. The role of ID in proteins of these

trafficking routes had not been previously investigated. We hypothesized that IDP/IDPRs are abundant and mediate many processes in these trafficking routes. Additionally, we hypothesized that IDP/IDPRs prevalence in these trafficking routes may explain some of the functional and evolutionary differences exhibited by these routes. Thus, we investigated the location and abundance of IDP/IDPRs and different functions they may mediate.

Last, we present the seminal work that provided the conducting thread of this dissertation work. We analyzed protein structures having large clusters of dehydrons[21] (i.e. backbone hydrogen bonds which are insufficiently protected from water) as markers of structurally unstable protein regions at the boundary between order and disorder. We hypothesized that these unstable regions could act as promoters of protein interactions that would aid stabilizing their structure.

As a whole, we provide a multi-scale and multi-organism assessment of disorder-mediated protein functions. This work contributes to understanding the functional modalities enabled by IDPs/IDRs, to elucidating the relationship between network and organismal complexity and ID, and to evidencing the evolutionary advantages ID offers.

Objectives

The overarching goal of this research is to identify functionally relevant protein features from sequence information. A particular focus will be devoted to identifying the role that intrinsically disordered regions may have in determining protein function by analyzing their sequence-function relationship in different biological systems.

The project is divided into the following three main objectives.

1. **Extract functional and structural features of proteins through the integration of computational protein sequence analysis, with the ultimate goal to inform functional experimental assays.**

   Specific Aim 1.1**:** Integrate sequence analysis tools with literature based information to predict the effect of mutations in the gene encoding alpha synuclein on the protein's aggregation propensity.

   Hypothesis: *specific regions in the gene encoding alpha synuclein alter the protein's aggregation.*

   We will test this hypothesis by providing a rational strategy to design different protein variants with different predicted aggregation propensities.

   Specific Aim 1.2: Create a functional profile of protein residues based on sequence analysis methods that allows the subfamily classification of the

racemase protein family, and guides experiments to modulate the substrate specificity of given subfamily members.

Hypothesis: *Specific protein residues in alanine racemase sequences alter the proteins' substrate specificity.*

We will test this hypothesis by using the functional residue profile information to design protein variants with different substrate specificity.

2. **Identify functional roles of intrinsic disorder in proteins from different biological systems and how intrinsic disorder affects cellular processes at the molecular level.**

Specific Aim 2.1: Perform a genome-wide analysis of intrinsic disorder and its relation to Gene Ontology functional classes in the model organism *Arabidopsis thaliana*.

Hypothesis: *Intrinsic disorder provides a mechanism to increase Arabidopsis thaliana's ability to adapt to the environment.*

We will test this hypothesis by assessing the level of intrinsic disorder present in *A. thaliana*'s proteins, focusing on the biological functions they perform. Additionally, we will compare the disorder of common functional classes of Arabidopsis and human proteins.

Specific Aim 2.2: Identify the roles of intrinsic disorder in proteins involved the main vesicle trafficking routes in human and yeast cells.

Hypothesis: *Intrinsic disorder may be responsible for some of the functional and evolutionary differences present in the main vesicle trafficking routes*.

We will test this hypothesis by performing an analysis of intrinsic disorder of the proteins belonging to the main vesicle trafficking routes in human and yeast.

3. **Identify the relationship between dehydrated protein regions, protein-protein interactions and disordered regions.**

Specific Aim 3.1: Identify clusters of dehydrated sites (e.g., sites which are not sufficiently protected from water) in soluble proteins and their functional implications.

Hypothesis: *Large clusters of packing defects in soluble proteins constitute structural singularities that are intermediate between ordered and disordered structures and mediate protein-protein interactions.*

Chapter 1

Background

Current projects for massive characterization of proteomes are generating an avalanche of protein sequences with unknown functions. Strategies to assess protein function based on computational methods have proven an effective approach for associating functional information to their sequences. This chapter discusses the underlying concepts behind the approaches we implemented in our own work to extract functional features from primary sequences.

In the first section, we discuss sequence analysis methods for the prediction of functional regions in proteins. We will particularly focus on methods to identify differentially conserved residues. In the second section, we introduce intrinsic protein disorder as another functional feature extracted from sequences. We

present a functional and structural classification of ID to demonstrate its role in modulating protein function and to emphasize the current need for a better understanding of this phenomenon. Additionally, we discuss the milestones associated with the development of methods to identify disordered regions, the current status of the field and future challenges.

1.1. Computational prediction of functional regions in proteins

The first fully sequenced genome was published in 1995: it was meningitis causing bacteria *Haemophilus influenzae*[22]. The number of sequenced genomes has since increased exponentially. Only 17 years later there are around 2000 completely sequenced genomes according to the statistics of GOLD[23] (Genomes OnLine Database)[23], which collects all genome and metagenome sequencing projects around the world. As more and more organisms are successfully sequenced, massive amounts of data are produced. To this date, the manually annotated section of UniProtKB[24] resource (Universal Protein Resource Knowledgebase, Release 2012_07), the UniProtKB/Swiss-Prot, has 536,789 non-redundant protein sequences, while UniProtKB/TrEMBL (unreviewed, automatically annotated) contains 23,165,610 sequences. Although these genomic data encode for many biological key features, they cannot be easily translated into meaningful biological information.

The current post-genomic era faces a titanic task, one that will probably span many decades: decipher genomic sequences to obtain functionally relevant

information. The main challenge of this endeavor is that as the number of protein sequences grows exponentially; it is impossible to experimentally derive their biological functions, let alone identifying the particular protein regions responsible for such functions. Consequently, the gap between the number of known protein sequences and proteins whose functions have been experimentally characterized is constantly growing[25].

Experimental approaches to determine functionally important sites are expensive and time-consuming hence often bound to small-scale studies. One example is site-directed mutagenesis, in which residues are replaced in a systematic way and the effect of the mutation is assessed (e.g. in binding to other proteins or changes in the protein activity)[1]. This type of experiments can be implemented only if some previous information on sequence specificity is available. Computational methods, on the other hand, can analyze copious amounts of data in a more resource-friendly manner and without the need of previously known functional information. In addition, analysis of computationally derived results can provide candidate functional sites that can then be tested by experimental techniques.

1.1.1. Prediction of functional sites

Functional residues are defined as residues required by the protein to perform its molecular function or biological role, and which cannot be changed (except for mutations to "compatible" amino acids) without affecting those functions[26]. Consequently, determining functional residues is a crucial step to

understand the protein's molecular mechanism. Once these residues are identified, it is possible to devise ways to address mutations at these positions (e.g. to revert pathologies derived from mutations) or to develop new functions (e.g. biotechnology). As previously discussed, computational approaches for predicting functional sites and features based on sequence and/or 3D structure information represent an alternative strategy to overcome the difficulties of determining them experimentally.

The majority of these computational methods part from the evolutionary principle that important sites tend to be preserved, meaning that amino acids in those positions can cannot be changed without altering their specific function. From the user perspective, the distinction of prediction methods based solely on sequence from those that incorporate structural information depends on the available information about the target protein. However, even if structural genomics projects are providing new structures at a growing rate, the ratio of known sequences to known structures is still low: the PDB[27] (Protein Data Bank), which collects all experimentally determined protein structures has 77,740 entries (August 23-2012). The limitations of 3D structure determination, together with the abundance of sequence information available, have ultimately made the development and application of sequence-based methods widely diffused. The general strategy of these methods is usually based on exploiting the power of sequence alignment and clustering[5].

1.1.2. Protein sequence alignments and functional sites

The most common and straightforward strategy for function prediction methods is based on sequence homology[28]. From an evolutionary perspective, two proteins are considered homologous if they share a common ancestor[29]. Accordingly, homologous proteins have similar sequences and tend to perform similar functions. These proteins can be grouped and their sequences aligned in multiple sequence alignments (MSAs) from which evolutionary information can be obtained. A MSA is usually represented as a matrix in which rows represent the proteins, and columns represent equivalent resides, and they are often referred to as *positions* of the MSA (Figure 1-1). MSAs provide a representation of the amino acid changes (due to structural or functional requirements) allowed by evolution at each position, thereby offering a rich source of structural and functional information[30].

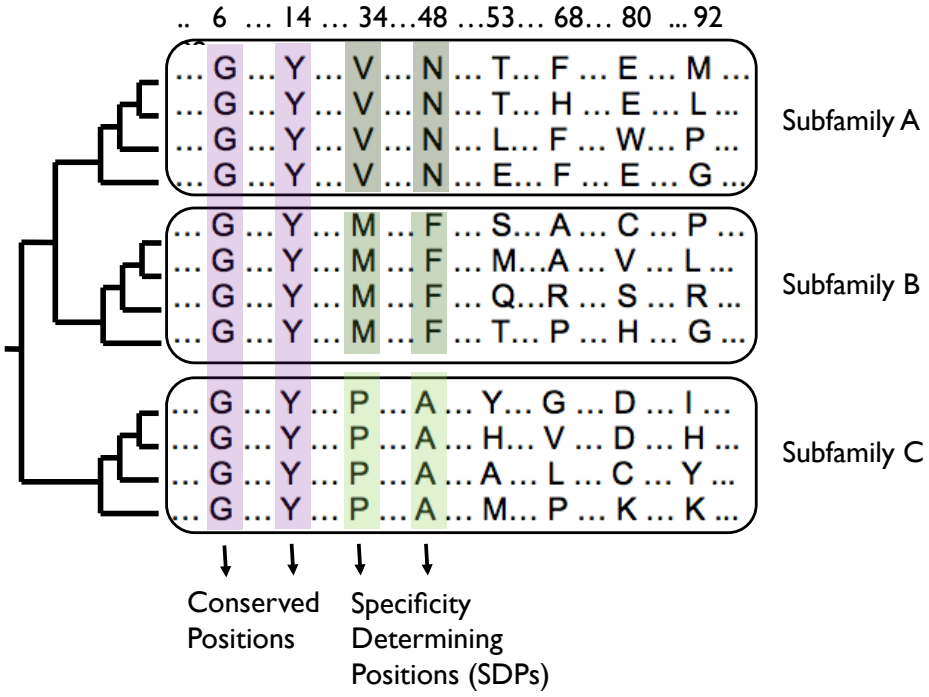

Figure 1-1. Representation of a portion of a multiple sequence alignment of the sequences of 12 homologous proteins. Rows represent proteins, and columns (positions) represent equivalent residues. Three subfamilies are defined in this alignment. Fully conserved positions (purple) are important for the whole protein set. Positions with a subfamily-dependent conservation pattern (green) are related to functional specificity.

1.1.2.1. Conserved positions

As previously mentioned, functionally active residues tend to be conserved among homologous proteins. Hence, fully conserved positions in a MSA are the first and most obvious pattern to explore. They most likely represent important residues for the protein's function and structure, since no changes have been allowed in those positions throughout evolution. Historically the first detected sites of

functionality[31], fully conserved positions can capture all different functional types of sites, including catalytic, ligand-binding, protein-protein-interaction and nucleic acid binding. However, not all conserved positions are necessarily functionally important; positions can also be conserved due to structural constraints. In general, these two types of conserved positions can be distinguished by the amino acid type, since it is known that, when conserved, some tend have structural roles, while others are usually involved in binding sites[32,33].

Locating conserved positions is not a trivial task. Methods vary from very simplistic approaches based on percentage of identity of the amino acids in the position (column), or the conservation according to the amino acid physicochemical properties, to more elaborated metrics such as Shannon's entropy, or the variance with respect to the mean amino acid distribution in the whole alignment[34]. Consequently, different methods may yield different results, also because sequence conservation is not perfect, for example, due to *conservative* substitutions for amino acids with similar physicochemical properties. Additionally, if the MSA contains highly redundant sequences, there will be many conserved residues, which in turn may be less indicative of functionality. Some methods, such as Conseq/Consurf server[35] – one of the most popular implementations for detecting conserved positions – incorporates the sequence phylogeny to avoid potential artifacts introduced by peculiarities or by the uneven distribution of sequences in the MSA (e.g. in case of highly similar sequences).

Residues forming functional sites typically cluster in the 3D structure of a protein, as they are collectively involved in the protein function. In addition, they tend to locate in the protein surface to be accessible to interacting molecules. Thus, if the 3D structure of homologous protein is available, it may be used to map the conserved positions onto its structure. By assessing their spatial clustering and surface exposure, positions that do not satisfy these constrains can be discarded.

Not all functionally relevant sites are fully conserved; some positions may show a distinct amino acid distribution in the MSA. Sequence profiles extracted from MSAs can be used to detect both fully conserved positions and the distribution of amino acids within positions. These profiles usually encode each position using a vector of 20 components that represents the fraction of each one of the 20 amino acids (where a fully conserved position would be coded as '1' in the corresponding position and '0' in every other position).

1.1.2.2. Family-dependent conserved positions (SDPs)

Proteins in a MSA can sometimes be partitioned into subgroups, and then the original concept of conservation can be expanded to consider also those positions with a differential conservation pattern among the different subgroups within the given MSA. Such positions may be conserved in a given subgroup but not in another, or the conserved amino acid might be different among the subgroups. The fact that these positions are conserved makes them functionally important, while the fact that the amino acid is different in each group indicates that this relevance is group-

specific with respect to the criteria used to define the groups. Accordingly, fully conserved positions (Figure 1-1, highlighted in purple) are important for the whole family of proteins, and positions with subgroup-dependent conservation (Figure 1-1, highlighted in green) are related to functional specificity. The group (also known as family) partition can be done according to different criteria, such as phylogenetic (if the groups evolved independently) or functional (if the groups have slightly different functions). When the subgroups (subfamilies) are defined using functional criteria, the positions relate to the functional differences (i.e. functional specificity) of each of the subfamilies, and they are known as *specificity-determining positions* (SDPs)[36,37,38,39], or *tree determinants*[40,41,42].

A number of approaches have been developed to identify SDPs. The evolutionary trace (ET)[43] method, for example, was one of the pioneer approaches and has served as starting point for the development of a number of other methods. ET hierarchically partitions the MSA into subfamilies following its phylogenetic tree and looks for the conserved positions that become apparent in each partition. Fully conserved positions appear in the whole MSA (root partition), while the first partition will reveal those positions which are differentially conserved in the two main subfamilies of the MSA, until the last partition of the tree is reached (having one protein per subfamily), in which all positions are conserved. A rank is assigned to each partition according to the position at which it becomes conserved. Positions with higher ranks are predicted to be functionally important, and experimental evidence showed that they are generally related to functional and binding sites[26].

Another set of methods for detecting SDPs is based on the vectorial representation of the MSA in a high dimensional space[1]. Each protein is represented as a vector based on its amino acid sequence and a dimensionality-reduction method (such as principal component analysis (PCA) or multiple correspondence analysis (MCA) is applied. These transformations result in equivalent spaces of a reduced dimension, which, while preserving most of the information, allow the identification of the main sources of variability in the MSA. Thus, vectors representing proteins with high sequence similarity will be clustered in the same regions of the "sequence space", allowing for the identification of the internal organization in subfamilies (subgroups)[1]. A similar vectorial transformation for the individual positions results in a residue space where SDPs are located in the same regions in space where the clusters representing the subfamilies are. In Chapter 2, we report a study case in which we used the S3Det[44,45] method to search for SDPs in the racemase family of proteins.

A third set of methods compare the mutational behavior of a position, or of multiple positions, to that of the whole alignment[1]. These methods are based on the assumption that the mutational pattern of the positions with a family-dependent conservation pattern should imitate the one of the whole alignment, given that this is the expected behavior for positions that are conserved differently in different subfamilies. In one of such methods, Xdet[26], also used in our analysis of the racemase protein family (Chapter 2), a matrix containing physicochemical similarities represents the mutational behavior for all pairs of amino acids at a given

position. The mutational behavior of the whole alignment is encoded by an equivalent matrix containing the overall similarities for all pairs of proteins. The comparison of these matrices yields a score for the position of the MSA, where highest scores are selected as predicted SDPs.

SDPs identification methods can also benefit from 3D information when available: in most cases it is used to evaluate the clustering and surface accessibility of the predicted positions[44,26].

Conservation-based predicted functional sites (either fully conserved or SDPs) cannot be further classified according to their specific role in the protein[1]. Methods for identifying these positions are based on the assumption that residues are conserved during evolution if they are functionally important, regardless of the function they perform. Thus, these sites can be identified as "important" for the protein, but nothing can be inferred regarding their role in protein interaction, in ligand binding, or in the catalytic activity[1]. Some methods, however, have been tailored to identify specific types of functional features, or to predict protein interaction (or binding sites).

1.1.3. Structure-based methods

The most straightforward and simple structure-based approach incorporates predictions from sequence-based methods to the available 3D structure information. By mapping the predicted positions to the 3D structure, it is possible to verify if they satisfy the expected structural conditions, i.e., if the sites are clustered or solvent-accessible. Several methods that identify clusters of conserved residues in the protein's surface assess residue conservation from the sequence distribution of the MSA as described in Section 1.1.2.1[46,35]. Examples include methods based on searching for conserved apolar residues clustering in the protein's surface[47], identifying surface regions that share the same phylogeny of the protein family[48].

An increasing number of structure-based methods for predicting functional sites are based on the general idea of locating "unstable" or "unusual" regions in the protein surface under the premise that interaction with partners (ligand or protein) restores their stability. Some approaches, for example, try to determine surface patches of residues with unusual physicochemical properties[49] or with specific binding thermodynamics[50] and patches defined according to specific electrostatic and solvation scoring[51], or solvent accessibility[52]. In Chapter 5, we present our own strategy based on the clustering unstable (due to water exposure) backbone hydrogen bonds in protein structures to detect regions involved in protein associations.

1.2. Protein disorder, not so messy after all

This section introduces disordered regions of proteins as a particular type of functional regions extracted from protein sequence. It was not long ago that Dunker *et al.* [53,54,55], and Wright and Dyson[56], among others, were striving to convince the community of the existence of proteins that, defying the classical structure-function relationship paradigm, lacked 3D structure and yet were fully functional. In the last decade, however, structural biology has seen groundbreaking advances in the relatively young field of protein intrinsic disorder, or as some like to call it, *unstructural biology*.

Intrinsically disordered (or unstructured) proteins (IDPs) were initially regarded to be anecdotic and isolated cases, such as when Sedzik and Kirschner first questioned myelin's ability to crystallize[57]. Despite the fact that Linus Pauling in 1940's suggested that disordered regions' flexibility could be an advantage for antibody creation[58], IDPs were generally not considered to perform important cellular functions. Research on IDPs was delayed not only by these misconceptions about their abundance and functional relevance but also by the fact that structural and molecular biology techniques were designed based on the paradigm of the structure-function relationship of ordered proteins.

Intrinsic disorder (ID) characterization has greatly benefited from the advantages that computational predictions offer over experimental techniques. Once the first IDPs and intrinsically disordered protein regions (IDRs) were

experimentally determined, computational tools allowed the identification and characterization of other IDPs and IDRs, providing crucial information about individual proteins, groups of proteins and, most importantly, entire proteomes[55, 54]. The first protein disorder predictors were based on the special amino-acid composition of IDPs. Williams *et al*. provided the first indication that IDPs' amino acid compositions differ from those of structured proteins by noticing the abnormally high charge/hydrophobic ratio in IDPs[59]. Ordered and disordered regions identified by X-ray crystallography, NMR and circular dichroism (CD) were subsequently used as input to develop neural networks to predict disorder from amino acid sequences[53,54,60]. The analysis of amino acid preferences of disordered protein segments revealed that IDPs' low overall hydrophobicity and large net charge accounts for their inability to fold into well-defined structures.

In addition to providing evidence of the existence and abundance of IDPs and IDRs, prediction methods also helped shifting the traditional structure-function paradigm by demonstrating that proteins can lack structure and yet be functional[56]. Other important milestones of the ID field include the development of the first database of experimentally determined disordered regions (the Database of Protein Disorder, Disprot[61]) and evidence that ID is involved in diseases such as cancer and neurodegenerative disorders[16,15]. In the last decades, researchers have discovered scattered evidences of IDPs playing important biological roles. However, only in the past few years the discovery and characterization of IDPs has flourished becoming one of the fastest growing areas of protein science[8].

A general consensus regarding IDPs' definition is still lacking. Particularly, it is not clear if IDPs are defined as proteins in which the entire sequence or significant segments of the sequence are disordered. Both, proteins completely lacking 3D structure and proteins containing only a few IDRs, are commonly referred to as IDPs. According to DisProt, an IDP is a "protein that contains at least one experimentally determined disordered region"[61]. DisProt collects and curates all structural and functional information available on experimentally identified IDPs and IDRs. On its current release (Release 6, July 01-2012), it contains 667 proteins and 1,467 protein regions. Thus, DisProt provides reliable data (such as IDP-related protein functions) that can be used in bioinformatic projects, including the development and testing of prediction methods. The Critical Assessment of protein Structure Prediction (CASP) experiments – experiments aimed at establishing the current state of the art in protein structure prediction – have recently included the assessment of protein disorder prediction methods[62,63,64,65,66].

There have been several attempts to measure the abundance of IDPs in different genomes, yet the results are not in agreement. Conservative estimates suggest that 5-15% of proteins have completely disordered sequences, where about 30-50% of the proteins have at least one long (of at least 30 consecutive amino acids) disordered region[11]. In mammals, for instance, 75% of the signaling proteins are predicted to contain long disordered regions[8]. According to recent estimates, 40% of all human proteins contain at least one long disordered region, and about 25% are predicted to be completely disordered in their sequence[67]. In general,

studies claim that the overall trend on the amount of disorder is higher for eukaryotes than for archea and eubacteria, with multicellular eukaryotes having much more predicted disorder than unicellular eukaryotes[68]. The fact that the amount of IDPs/IDRs increases with organismal complexity could be explained by the observation that disorder is abundant in processes intuitively related to complexity, such as those involved in the coordination of various organelles of eukaryotes (e.g. signaling)[11,69,70]. This is also supported by the notion that ID plays a key role in protein-protein interactions[71,72,73], especially in moonlighting, when proteins bind to different partners and perform different functions[74]. The relationship between organismal complexity and ID will addressed in detail in Chapter 3, in which we analyze the function of the most disordered proteins in *A. thaliana*.

According to several genome-wide computational studies, IDPs functional roles include: (i) transcription and regulation, (ii) signal transduction and cell-cycle regulation, (iii) functioning of nucleic acid containing organelle, (iv) mRNA processing and splicing and (v) cystoskeleton organization[75,11,10]. The molecular functions which allow IDPs to perform those functional roles are mainly associated with molecular recognition, chaperone activity, RNA and DNA binding.

The revolutionary shift in the structure-function paradigm introduced by the notion of protein disorder has brought new challenges: devising effective strategies to predict and eventually experimentally test ID in living cells with the ultimate goal

of characterizing its role in protein function. In the next section, we will discuss current attempts to classify ID according to its functional and structural functions.

1.2.1. Structural and functional classification of protein disorder

Criteria to classify IDPs from their structural and functional features have recently emerged[76,19]. From a biophysical perspective, ID is considered a structural state of proteins that corresponds to an ensemble of rapidly interconverting conformations. Thus, ID is usually classified using a continuum of structural states spanning from folded structures – which might also have disordered regions – or "compact disorder" to fully disordered states or "extended disorder"[77,78,79]. This structural spectrum is subdivided into three main categories: i) proteins that may fold upon binding, thereby undergoing a disorder to order transition, ii) proteins that remain unfolded even in their bound state, in which the ability to maintain a highly flexible conformation enables interaction (e.g., function), and iii) proteins with more structural constraints that might adopt a molten globule conformation, but that are still regarded as disordered[76]. Following this classification, IDPs may have a wide range of compactness (i.e. ratio of accessible surface area of a protein to that of the ideal sphere of the same volume[80]), secondary structure content and number of tertiary contacts[19]. According to Uversky and Dunker, IDPs constitute the "fourth tribe" of proteins to be added to existing main traditional classes of fibrous, globular and membrane proteins[19,81].

Protein disorder also affects protein function. Disordered stretches can serve as *flexible linkers* that function as spacers allowing conformational changes altering the relative orientation of different structural domains within a protein[82]. The function of these entropic chains relies on their ability to rapidly fluctuate among alternative states with similar conformational energies. Acting as entropic springs, bristles, spacers, linkers, they generate force against structural changes or influence the orientation or localization of the attached domains[83]. By definition, the functions of flexible linkers cannot be fulfilled by rigid structures. Moreover, multidomain proteins cannot exist without disordered linkers that regulate distance and enable freedom in the orientational search. Examples of this flexible regions include a region of the protein titin (which varies from 180 residues to 2,174 residues, depending on the isoform considered), involved in maintaining the appropriate length of muscle fibers[84] and the entropic bristle that provides spacing in the cytoskeleton and the FG repeat region of nucleoporins[85,86]. In Chapter 4, we will discuss the function of these disorder-based linkers in the context of vesicle trafficking proteins.

Most often, disordered regions are sites of *molecular attachment* that may become ordered upon interaction with one or several binding partners and give rise to functional specificity[87,8,82]. IDRs involved in molecular recognition are typically involved in transient or permanent interactions with macromolecules or small ligands and they can be classified according to their function[88,89]: i) display sites and ii) chaperones, which are typically involved in transient interactions (often

weak and of limited specificity[18]), and iii) effectors, iv) scavengers and v) assemblers, which are involved in permanent interactions.

Display sites, such as phosphorylation, ubiquitination and proteolytic attack sites tend to occur in disordered segments of proteins[90]. Examples include the ubiquitination sites of securin and cyclin B[91] and the degradation of non-ubiquinated disordered proteins such as tau, casein[92,93], and p21[94] by the 20S-proteasome. Among *chaperones*, those assisting RNA folding seem to be specially enriched in disordered regions (e.g. nucleocapsid protein 7/9, ribosomal S12, the prion protein N-terminal domain)[89]. Disordered proteins acting as *effectors* bind and modify the activity of their partner enzyme. Examples of effector proteins are p21 and its homologue p27, which inhibit cyclin-dependent kinases[88] while at the same time promoting assembly of the cyclin-Cdk complex leading to Cdk activation[95].*Scavenger* proteins are typically used to store and/or neutralize small ligands. Casein, for example, prevents calcium precipitation in milk by sequestering small clusters of calcium phosphate[88]. Structural flexibility is also crucial in *assembler* proteins (or domains), which assemble, stabilize and regulate multi-protein complexes. These proteins bring multiple partners together so they can interact, and play important roles in many cellular functions including the assembly of cytoskeleton, ribosome, and the chromatin[82]. In many cases, assembler proteins rely on the fly-casting mechanism for protein binding, where long disordered regions rely on their bigger capture radius to efficiently span the environment for their binding partners[96]. In Chapter 4, we report examples of disordered proteins

involved in vesicle trafficking that use fly-casting to assemble large macromolecular complexes.

The classification reported above allows to systematically group the functional roles of IDPs. However, a single protein or even protein region can also combine multiple functions. IDPs' functions complement those of structured proteins having more topological constraints: certain complexes, for example, cannot be assembled from rigid components.

The conformational behavior that IDPs exhibit upon molecular recognition of their binding partners involves many scenarios. They usually undergo coupled binding and folding while binding the partner[8,97]. Alternatively, binding of disordered regions may also depend on conformational selection, which seems to occur via "pre-formed" structural elements that serve as initial contact points[58,98]. In many cases, however, disorder mediated protein-protein interactions actually involve a combination of coupled binding and folding with conformational selection[99,8]. The structural transitions of highly flexible regions may often depend on their binding partners, which allows them to interact specifically with structurally unrelated partners in a process called binding promiscuity or one-to-many signaling[100,101,102]. The flexible C-terminal domain of p53, for example, interacts with more than forty partners. The p53 binding region may adopt very different types of secondary structures upon binding to unrelated partners: from an α-helix (when bound to S100ββ) to a β-strand (when bound to situin) to coil (when

bound to CBP and cyclin A2). Moreover, different residues are involved in these interactions[102].

Short segments in disordered regions (usually l0 to 15 residues) that modulate molecular recognition have been extensively studied and designated as molecular recognition features (MoRFs)[103,104,105], eukaryotic linear motifs (ELMs)[106], short linear motifs (SLiMs)[32], and ANCHOR regions[107]. ELMs and SLiMs are based on sequence motif identification, while MoRFs and ANCHOR regions are extracted directly from disorder predictors.

Molecular recognition might also be mediated by longer disordered regions (about 20-30 residues) that correspond to whole disordered domains[108]. According to Tompa *et al*., these regions actually satisfy the classical definition of protein domain: i) they are structurally and functionally independent elements of the protein, ii) their sequence is evolutionary conserved (hence it can be recognized by homology) and iii) they possess at least one specific biological function. There are several existing protein domains reported in the Pfam[109] database belonging to this category, while many long disordered regions correspond to domains that are likely not yet represented in Pfam.

Disordered domains and recognition motifs do not only differ in the length of their sequence. Linear motifs constitute effective evolutionary switches that can be randomly turned on and off by point mutations (especially if several motifs are required for the recognition)[110,111]. Thus, their occurrence within different contexts

can be regarded as an example of evolutionary convergence[69]. Disordered domains on the other hand, represent functional units that are spread in the genome by inheritance, which suggests evolutionary divergence[112].

1.2.2. Protein disorder from an evolutionary perspective

As structure is closely related to function, it is also subject to evolutionary pressure. Protein evolution is usually empirically described through the comparison of homologous proteins. Dayhoff *et al*. proposed a model to evaluate evolution of proteins based on assessing the frequency with which different amino acids occur in a given position among the different homologs[113]. According to this model, only point mutations having neutral or positive effect on the protein function will be selected during evolution.

The evolutionary process of IDPs is still not fully understood. IPDs were found to generally evolve at a significantly faster rate than structured proteins[114]. However, other studies indicate that individual protein families maintained the features of disordered regions even if subject to rapid evolution[115,116,117]. Specifically, disordered regions were maintained in length and flexibility even in cases where the amino acid sequences where not maintained[76]. Chen *et al*. found several examples of protein domains and families with conserved disorder [118], and even assembled a database with such conserved regions[119].

According to Brown *et al*., however, since some of these studies were based on models biased towards evolutionary changes found in structured proteins, it was not surprising that disordered regions were lost[114]. To overcome this limitation, they developed and compared models of evolution for IDPs and structured proteins, and noticed that IDPs have unique patterns of amino acid substitutions when compared to their structured counterparts[120]. Szalkowski and Anisimova also showed that there are significant differences in the evolution of disordered and structured proteins[121]. They observed that disorder promoting amino acids are more conserved in IDRs than in structured regions, suggesting that not only the amino acid composition, but also the specific sequence is important for function. Interestingly, they also reported that in almost one third of their dataset, IDRs evolved more slowly than the structured segments of the proteins. Additionally, other studies reported that the sequence conservation of disordered Pfam protein domains was similar to that of structured Pfam domains[119,108].

Identifying structural and biological factors that influence the wide range of evolutionary rates of IDRs is an ongoing challenge. We believe that it is crucial to distinguish between the two main functions of ID: flexible linkers and molecular recognition. The only constraints of entropic chains appear to be maintaining the flexibility and the length of disordered segments, whereas in disordered regions mediating molecular recognition, sites responsible for interactions are likely to be under selective pressure as the specific sequence is strictly linked to the protein's

function. Protein modification sites, for instance, may be constrained because mutations in these sites could have a deleterious effect on signaling events[122].

Recently, Mosca *et al*. showed that the conservation of disorder facilitates the change of interacting partners during evolution[123]. Furthermore, different studies have suggested that nature uses protein disorder to adapt to different environments[124]. Archea proteins are relatively rich in disorder, which is thought to help them accommodate to hostile environments[125]. Similarly, organisms with high tolerance for mutations tend to be enriched in disorder. This is the case of *Deinococcus radiodurans*, a bacterium capable of surviving high doses of radiation[126]. In general, it appears that harsh environmental conditions tend to favor increased disorder content. . Disorder could be therefore regarded as a buffer against deleterious mutations[124]. The conservation of flexibility and function of disordered regions, seem to suggest the existence of a selective pressure to maintain disorder throughout evolution. This conservation seems to confirm the relevance of IDPs' functional roles[127,128].

In summary, there may be multiple interpretations on the evolution of intrinsically disordered regions. Furthermore, it is still unclear to what extent disorder needs to be conserved in order to preserve function. It is clear, however, that ID remains a mechanism nature has evolved to contribute innovation throughout evolution.

1.2.3. What's all the fuzz about disorder?

As discussed above, selective pressure to maintain the disordered nature of certain protein regions results from the inherent properties and advantageous phenotype that these regions provide to the protein function. The presence of IDRs is thought to confer many functional advantages when binding to their partners. The ability of IDPs to fold upon binding to the partner decouples the specificity from the binding strength, allowing for high-specificity and low-affinity interactions[129]. The binding specificity is mainly determined by the size and complementarity of the binding interface[129]. When the IDP region folds upon binding, there is an entropic penalty on the free energy of binding because the IDP region (previously unbound and free to move in solution exploring diverse conformations) is now fixed in a bound conformation[77]. The loss of entropy may reduce the binding energy thus producing a weaker binding[129]. By uncoupling specificity from binding strength, IDPs can increase the speed of interaction and provide adaptability, conferring IDRs the ability to bind to different partners[130,101,74]. IDPs offer a large and flexible interaction surface area which makes them very suitable for mediating signaling cascades where specific interactions with fast association/dissociation rates are required[131,132]. In addition, the availability of molecular recognition features in long disordered segments may enable remote initial interactions via fly-casting to initially search the surrounding environment for the different partners in a very rapid and efficient way[18]. Finally, the conformational variability of IDRs allows the binding surfaces to adjust to a range of diverse partners (known as binding

promiscuity or moonlighting[74]). As a result, this context-dependent folding activates or inhibits signaling processes that can have completely orthogonal outcomes. Flexible regions of IDPs may also facilitate access to enzymes and effectors that mediate post-translational modifications and recognize the post-translational code. Specific post-translational modifications allow combinatorial regulation and the use of the same protein in multiple biological processes[7].

The features of disordered regions of proteins described above may explain why proteins containing disordered regions are highly involved in a number of protein-protein interactions. Several studies have proposed that hub proteins (i.e. highly connected proteins) use disordered regions to bind to multiple partners, which results a single protein mediating different signals depending on the specific binding partner (one-to-many signaling)[133, 134,135,]. The abundance of IDPs as hubs has been reported in the protein-protein interaction networks of different eukaryotes including yeast, *C. elegans,* Drosophila and Human[136,137,138,139]. While generally enriched in disordered regions, most hubs in protein interaction networks contain a mixture of structured domains and long disordered segments and interact with their partners with both their structured and their disordered regions[77,140,141]. For structured hubs (or structured domains in hubs), it was also proposed that it is the disordered regions in their binding partners that mediate the interaction (many-to-one signaling)[133,142]. Finally, it is important to note that while IDP hubs are known to interact with multiple partners, many of the interactions formed may be

short lived, mutually exclusive or dependent on cellular localization and cell cycle. This promiscuity can sometimes hamper the identification of this interactions[7].

Recent studies showed that some IDPs do not fold (or fold only partially) even when they are present in their bound state, forming dynamic or *fuzzy* complexes[19,18,143]. Thus, the concept of protein disorder can be extended from individual proteins to protein complexes. Tompa and Fuxreiter showed several examples demonstrating that disorder in the bound state or *fuzziness,* can range from static to dynamic, and that it may involve the whole protein, or only parts of it[143]. According to their analysis, fuzziness spans the whole spectrum between these two extremes: s*tatic disorder*, where an IDP region may adopt multiple (but stable) conformations representing the classical view of disorder, and *dynamic disorder*, where an IDP or an IDP region may continuously fluctuate between many states (i.e. conformational ensemble). Dynamic disorder can be further divided according to the degree and dynamics of disorder of the interacting segments. Static disorder is thought to follow the *polymorphic model*: in the bound form, this type of complexes can range from having a few to having many alternative stable structures. The heterogeneity of the bound conformations will likely results in different effects on the binding partner, hence it is thought to help amplifying the functional repertoire. In addition, recent studies have suggested that IDPs can bind the same partner using different regions and adopting different structural conformations, leading to different functional outcomes[74], which suggests that fuzziness mediates regulatory functions. In dynamic disorder, bound proteins can fluctuate in an ensemble of

dynamic conformations. This dynamic binding is different from the classical understanding of protein binding interactions, in which binding implies bringing proteins together, and fixing them spatially and temporally. Tompa and Fuxreiter also present a schematic classification of the functional roles that disordered regions may play in dynamic complexes: *i) clamp model*, where IDRs can increase the conformational freedom and adaptability of the two binding regions (in this case, the disordered segment serves as a linker between two ordered recognition regions[144,145,146]); *ii) flanking model*, where disordered regions leave space for other binding partners, post-translational modifications or prevent aggregation (frequently observed when IDPs bind through short binding motifs: IDRs flanking the interaction interface remain disordered while the binding interface becomes structured[90,147,89]); or iii) *random model*, which is the extreme case of disorder: the entire protein remains disordered in the bound state[143,148](ideal for transient interactions).

The formation of fuzzy complexes has been described for a whole range of interactions, including: IDPs interacting with structured proteins[149,150,151,152], with other IDPs[153,154,155] and with biological membranes[156,157]. A similar binding mode is expected to apply for IDPs interacting with nucleic acids and other macromolecules[129].

In summary, fuzziness can be functionally advantageous in protein-protein interactions. Similar to what was observed in IDPs, it may add adaptability,

versatility and reversibility to the interactions, aiding in protein-protein interactions regulation.

1.2.4. Methods to predict and evaluate protein disorder

The highly dynamic nature of IDPs' precludes the determination of a unique high-resolution structure. Consequently, experimental methods are needed to identify constraints on the ensemble of states sampled by an IDP at multiple time scales. The integration of both computational and experimental techniques allows characterizing the full spectrum of structural and dynamic features of IDPs.

Biophysical methods allow determining IDPs protein structure and dynamic behavior at different time-scales. These are crucial for the structural characterization of IDPs and for determining the relationship between the highly dynamic structure of IDPs and their biological functions. Biophysical methods provide information on many of the IDPs features including their shape, conformational stability, overall compactness, residual secondary structure, regions of enhanced or restricted mobility, transient-long-range contacts[19,158].

The presence of disordered regions was first observed with X-ray crystallography experiments. In these experiments, amino acid loops known to be required for function were occasionally missing from high-resolution protein structures[159,160]. The high flexibility of the atoms in those "loopy" regions leads to non-coherent X-ray scattering, making them invisible. Thus, X-ray crystallography designated regions with missing electron density as disordered regions.

Results obtained with NMR spectroscopy also showed that some proteins with known biological functions did not have a stable, well-defined structure in solution[59]. This technique is based on the information provided by the molecule atomic nuclei and their local environments, and it is the best tool for providing high-resolution structural information on IPDs in solution. NMR spectroscopy however, presents some technical limitations, such as the limited protein size, lack of spectral dispersion due to similar environments for different residues, and increased redundancy due to the presence of tandem repeats (often found in disordered proteins). Additionally, this technique cannot provide information about the overall size and shape of IDPs. In-cell NMR spectroscopy, on the other hand, is used to characterize IDPs in their natural environments (i.e. within cells). This approach allows investigating the hypothesis that predicted disordered proteins are forced to adopt a 3D structure if present in the crowded cellular environment. This method has been applied to both bacterial and eukaryotic cells[161,162] and it has been widely exploited recently for studying ID in motor proteins such as kinesin and dynein[58]. Finally, heteronuclear multidimensional NMR spectroscopy can be especially useful for the direct measurement of IDRs' mobility[163]. It can also supply information on the extent to which IDRs engage in the formation of transient secondary structure elements[164].

Other techniques, such as Fourier-transform infrared spectroscopy and deep-UV resonance Raman spectroscopy, for example, can be used to determine the presence or absence of secondary structure elements. Likewise, circular dichroism,

Raman optical activity and optical rotary dispersion measurements, can be used to evaluate protein's secondary and tertiary structure.

Single-molecule studies can be very effective to describe IDP structure, since they allow to observe transient intermediates and both static and dynamic heterogeneity of structure[18]. Kodera *et al*. proposed a technique allowing the observation of both static and dynamic heterogeneity in IDP structure without ensemble averaging[165]. This novel technique, called high-speed atomic force microscopy, allows visualizing conformational transitions in structural disorder. It was used to witness the dynamic behavior of myosin V molecules translocating along actin filaments[165]. This experiment provides direct evidence of dynamic molecular behavior, leading to a comprehensive understanding of the motor mechanism. This approach seems very promising to study the structure and dynamics of IDPs in action.

The experimental identification of disordered regions is often convoluted and presents some limitations. Additionally, it can only be applied to small-scale studies involving few proteins. An alternative approach to predict disordered region is provided by computational methods.

1.2.4.1. Theoretical basis of computational methods to predict proteins' intrinsic disorder

Computational methods have quickly become a particularly valuable tool, especially because they can efficiently manage data from large-scale genome

sequencing projects. Their popularity has also been boosted by the fact that CASP (Critical Assessment of Structure Prediction) experiments have included disorder prediction to their tasks since 2002[62,63,64,65,66]. These techniques are based on the premise that amino acid sequence should encode for non-folding just as it encodes for protein folding. Predictors can be usually classified into three main categories: i) propensity based predictors, ii) machine learning algorithms, and iii) algorithms based on interresidue contacts. This division is not mutually exclusive, and some predictors may use more than one approach. In addition, several predictors can be combined to give rise to a forth category: iv) metapredictors.

Propensity-based predictors

These methods are based on the amino acid composition of the sequence. Amino acids are classified according to Dunker and Romero's assessment of the abundance of residues in disordered vs. ordered protein segments (also used in other disorder prediction approaches)[70,166]. Dunker *et al.* showed that IDPs are significantly depleted in order-promoting residues, which include bulky hydrophobic amino acids (Ile, Leu and Val) and aromatic amino acids (Trp, Tyr, Phe) that would normally form the hydrophobic core of the protein along with Cys and Asn. Conversely, disorder-promoting residues include Ala, Arg, Gly, Gln, Ser, Pro, Glu and Lys.

Methods based on these amino acid propensities are simple and easy to implement, but they rely solely on the protein's amino acid composition. Another

approach is to use the biased amino acid composition of IDRs: they have low overall mean hydropathy (i.e. sum of the hydropathies of all residues divided by the number of residues of the chain) and high mean net charge (i.e. net charge at 7.0 pH divided by the total number of residues). The rationale behind this approach is that high net charge leads to charge-charge repulsion, while low hydrophobicity implies a weak driving force for the formation of a compact structure[167]. Several methods use different measures based on charge and hydrophobicity. *Foldindex* for example, calculates the distribution of the mean charge and mean hydrophobicity for a predefined sequence window, providing a per-residue disorder prediction[168]. *Globplot* is based on the relative propensity of residues to be in an ordered/disordered state according to a predefined amino acid scale based on the difference in the probability for a given amino acid to be in a secondary structure or to be in random coil[169]. This approach identifies ordered domains thus eliminating segments that are predicted to be ordered, but that are too short to fold.

Machine learning algorithms

A very popular approach for disorder prediction is based on the use of standard machine learning techniques such as neural networks (NNs) and support vector machines (SVMs) to classify protein regions as ordered or disordered. These approaches are based on the assumption that sequence features calculated from a local sequence window can be used for a binary classification of order/disorder[167]. A great number of predictors use this approach; trained on datasets of order and

disorder, they use protein sequence as an input and provide a per-residue prediction of disorder.

The *PONDR* family of algorithms[60], for example, use methods based on NNs, while others use a combination of NNs and SVMs. The inputs of these predictors are sequence features (e.g. coordination number, hydropathy, net charge, and the fraction of the various amino acid groups) calculated within a sequence window. The training sets are different in the various algorithms (e.g. missing residues in X-ray structures, characterized on disordered regions, DisProt). They are especially useful for identifying regions that potentially undergo order-to-disorder transitions[105,104]. *DisEMBL* uses an ensemble of feed-forward networks that separately predict three kinds of disordered structures: i) residues within loop/coils, ii) residues in hot loops (with high B-factors, i.e. with high mobility) and iii) residues missing from X-ray structures[170]. This method is especially suitable for predicting short disordered regions.

Other machine learning algorithms are based on SVMs, which can be trained more efficiently and are less prone to overfitting than NNs[167]. They also penalize more miss-classification errors, which can be advantageous when using biased datasets. *DISOPRED2*, for example, is widely used and has a very low false positive rate[171]. The training set is based on amino acids missing from ~750 high resolution PDB structures, thus its performance is better on short segments belonging to globally structured proteins. The main distinguishing property of this predictor is the fact that it is not trained on amino acid composition measures, but it is directly

trained on the whole protein sequence. *SPRITZ* is implemented by a non-linear SVM based on multiple aligned sequences and consists of two separate predictors for long and short disorder regions[172].

Prediction methods based on interresidue contacts

Prediction methods based on structural and energetic features not relying on experimental data allow overcoming limitations associated with biased and insufficiently populated datasets. They are based on the assumption that disorder in proteins is a consequence of the lack or low level of interresidue contacts that are not able to compensate the large decrease in conformational entropy during folding[173]. Interresidue contacts are especially important in heavily interacting residue clusters, which are essential to stabilize the folded protein structure[174]. *FoldUNfold*, for example, calculates the expected average number of contacts per residue using an amino acid propensity scale that encodes the average number of contacts for the 20 amino acid residues in a dataset of globular proteins[175]. The average contact number of residues within a given distance in a protein structure depends on the mean packing density of the residues. Expected low packing density corresponds to disordered segments. *IUPred* is based on similar principle, but it has a more general approach for predicting protein non-folding: if a given residue does not form enough favorable contacts, it is assumed that it will not adopt a stable position in the 3D structure of the protein[176]. The direct estimation of the interaction energies uses only protein sequence whenever possible. The estimated energy for each amino acid depends on the nature of the amino acid and on the

composition of the neighboring amino acids, and it is summarized in a 20x20 energy predictor matrix. Residues with less favorable predicted energies are generally more likely to be disordered. The parameters of this method are defined based on a dataset of globular proteins, larger than any dataset of disordered proteins. Thus, this method is more stable than methods with a large number of parameters trained on a limited (and even ambiguous) disordered set of proteins. IUPred is capable of predicting both short and long disordered regions. Finally, *ANCHOR* is especially designed to predict biding regions located in disordered segments of a protein[107]. It identifies segments located in disordered regions that cannot form enough favorable interchain interactions, but are able to gain energetically by interacting with a globular partner protein. This method uses the same energy estimation method as IUPred.

Metapredictors

Using algorithms incorporating more than one prediction method is an alternative to overcome individual prediction bias and limitations. This approach, successfully used in many areas of structure prediction[177], allows reducing the noise of individual predictors. Combining the outputs of individual predictors, these methods provide predictions at the residue level as well as for the whole protein sequence. Their accuracies are usually higher than the individual predictors (10% increment)[178]. *MetaPrDos*[179], based on SVM, is a combination of per-residue individual predictors trained on a group of PDB-extracted proteins having regions with missing electron-density and less than 20% of sequence identity. *Meta-*

Disorder predictor (MD) is based on NN trained on proteins from PDB and DisProt and combines four complementing predictors[180]. *MeDor*, uses a whole battery of 13 individual algorithms, including disorder predictors, secondary structure prediction, hydrophobic cluster analysis[181].

1.2.4.2. Constraints and disadvantages of computational methods to predict proteins' intrinsic disorder

At present, several CASP experiments have evaluated the performance of different ID prediction methods. Predictions based on amino acid sequences are made in parallel to their structure determination. Once the structure of a given protein is completed, the results from the different prediction groups are compared. CASP experiments evaluate overall accuracy of the methods according to different measurements based on specificity and sensitivity. According to Monastyrskyy *et al*., although the number of participating disorder prediction groups has been increasing over the years, this does not correlate with an increase in the predictors' performance[6649,55,43,42]. Even metapredictors are still considered inaccurate as the best increments represent less than a 10% accuracy increase over individual predictors [178].

However, CASP comparisons of the different methods suffer from an underlying bias: the measurement of the methods' performance depends critically on both the type of disorder and evaluation criteria[167,141]. On the one hand,

predictors are usually trained focusing on specific types of disorder, which may lead them to perform poorly if confronted with other types of disorder. Different disorder predictions vary significantly in their performance on the DisProt database, with generally high sensitivity and low specificity values[167]. Thus, the choice of the disorder predictor is crucial. Some predictors, for example, classify sequences characterized as Charged Single α-Helices (CSAHs, i.e. sequences adopting stable helical conformation in water) as both IDPs and coiled. On the other hand, the predictors' performances also depend on the evaluation criteria used: while a given prediction method might be more focused on reducing its false positive rate, another method could penalize more the false negative rate.

Another limitation for further improvement in disorder prediction derives from the lack of appropriate datasets for training and testing the methods. The inaccuracy of disordered and ordered protein data is critical for the prediction methods' performance[167,130]. The dataset of experimentally determined IDPs and IDRs is still rather small, and it may contain misclassified segments. For example, in Disopred2, disordered residues for training are often identified as those appearing in sequence records but with coordinates missing from the electron density map. This can introduce errors, as missing coordinates can also arise due to artifacts in the crystallization process. False assignment of order may also occur as a consequence of stabilizing interactions by ligands or other macromolecules. In addition, the training protein segments are often too short to provide enough information[178]. Finally, experimental methods used to derive disordered segments

can also introduce bias on the length and location of the disordered regions[182,183]. Despite all these limitations, current accuracies of predictors are around 80% in terms of the averaged value of specificity and sensitivity according to recent CASP9[66].

The most important constraint for the further development of disorder predictions is the lack of conceptually novel methods. In our opinion, new models based on the nature of the different types of disorder are likely to enable significant progress in the field.

Last, the specific application determines the choice of the most appropriate predictor. Predictors especially developed to identify long stretches of disordered residues, for example, are more suited to conduct whole genome predictions, while ANCHOR can be used to evaluate the involvement of protein disorder in interactions[184].

1.2.5. Protein disorder, disease, and drug development

The vast implications IDPs have in many different cellular processes were described in the preceding sections. Not surprisingly, IDPs have recently been found to be tightly regulated[185,186], and, they have been often linked to diverse pathologies[14].

Huntingtin protein, responsible for Huntington disease, Tau protein in Alzheimer's disease and prion protein in prion disease are all well known IDPs[14,12]. So is alpha synuclein, whose misfolding and aggregation are apparently related to its

disordered nature and which is associated with the development of Parkinson's disease[13,14,15]. Interestingly, the high degree of association between ID and neurodegenerative diseases might be due not to the lack of structure of IDPs (which might lead them to aggregate) but mainly to IDPs and IDRs' functions as regulatory proteins and signal transducers. Therefore, mutations in IDPs/IDRs can impair the protein's function or expression and, in turn, this loss of function often leads to disease. More specifically, mutations in IDRs may impair their ability to properly identify binding partners. As previously mentioned, IDPs are often promiscuous interactors and serve as hubs in protein networks. The deletion of such heavily connected nodes is often lethal for the organism[187,39]. Failure to activate signaling cascades, for example, may lead to cancer. In fact, several cell-signaling and other cancer-associated proteins are enriched in disorder[16]. Tumor suppressor p53, c-myc proto-oncogene, FUS oncogene, Mdm2 oncoprotein and BRCA1, among others, all have long regions of disorder[8].

IDPs are also often found related to other diseases such as diabetes[79] and cardiovascular diseases[188]. Additionally, IDPs exhibit high dosage sensitivity[189,186]. Work on yeast (and inferred and validated in human and mouse) showed that overexpression of disordered proteins is harmful[189]. Apparently, when IDPs are present at high concentrations, their disordered regions are prone to make promiscuous molecular interactions, "which is the likely cause of pathology when genes are overexpressed"[189].

The association of IDPs with a variety of diseases has lead to recently consider them for drug design. IDPs' binding pockets resemble the active sites of enzymes. Thus, the binding partners of IDPs have been suggested as possible targets[190]. This is the mechanism of nutlin-mediated inhibition of p53-MDM2 interaction and reactivation of p53 pathway in cancer cells[191]. Lately, it has been proposed that IDPs can be targeted by small molecules. According to Dunker *et al*. IDPs' thermodynamic properties of binding, conformational flexibility and adaptability confer them significant advantages over structured proteins with respect to their potential to interact with diverse partners[8]. The general strategy to use IDPs as drug targets consists in developing small molecular drugs that induce the folding of the IDP, and enhance binding affinity by inducing conformational disorder in the target protein[7]. Thus, the identification of small molecules that specifically target interaction interfaces and structural transitions involving disordered segments could be a promising strategy for drug design[7]. For example, recent studies reported small molecules that shift the equilibrium of Myc-Max dimerization by binding to a disordered region in Myc, promoting its disordered state, and preventing its interaction with Max. This strategy decreases the possibility of Myc-Max overexpression in many cancers[17,192]. Last, a recent study suggested proline rich motifs in disordered regions as potential targets of immune related disorders[193].

In summary, it is now widely accepted that IDPs are part of essential cellular mechanisms, that ID might even be favored by evolution, and that IDPs also play a

key role in the development of human diseases. Now that it is demonstrated that the structure-function paradigm is only one side of the coin, it is time for “unstructural” biology to uncover the other side.

Chapter 2

Computational prediction of important regions in proteins of interest

2.1. Introduction

The overarching goal of this work is to integrate protein sequence analysis to extract functional and structural features of proteins with the ultimate goal to inform experimental assays. In this chapter, we will describe two case studies. In the first case study, we combined different sequence analysis tools with literature information to predict the effect of mutations in the gene encoding alpha synuclein (αsyn) on the protein's aggregation propensity. In the second study, we predicted a footprint of important protein residues based on members of the racemase protein family with different substrate specificity.

2.2. Alpha synuclein aggregation: predicting the sequence-structure relationship using rational design

α-synuclein (αsyn) is a highly conserved, largely intrinsically disordered protein that is abundant in neurons. αsyn accumulation and aggregation is linked to the development of a group of neurodegenerative diseases known as synucleinopathies[2]. Particularly, αsyn aggregation is the hallmark of Parkinson's disease (PD), the most prevalent neurodegenerative movement disorder[3,4]. The biophysical properties promoting protein aggregation into non-functional and usually toxic structures are similar to those mediating folding into native conformations[194]. However, the processes that govern αsyn aggregation are not fully understood. Discerning the molecular mechanisms that govern αsyn's misfolding and aggregation is fundamental for understanding and treating synucleinopathies[195]. However, studies to investigate aggregation determinants present a number of limitations. The transient and heterogeneous nature of αsyn's misfolding intermediates hampers the ability to obtain reliable data[196,197]. In addition, effectively monitoring and deciphering αsyn's misfolding and aggregation requires efficient approaches for detecting and quantifying αsyn solubility in living cells[195]. The techniques for monitoring αsyn aggregation in living cells are usually applied on few experimentally verified mutations that exhibit different aggregation propensities. However, the protein sequence and structural features of these protein variants have not been explored in a systematic manner to identify the determinants of protein aggregation. We propose an *in silico* strategy to investigate

the relationship between sequence-structure of αsyn. Our approach will allow predicting the protein regions that affect αsyn's aggregation propensity.

The aggregation behavior of polypeptides has been reported to largely depend on the intrinsic properties encoded in the sequence composition and in the primary structure[194,198]. Particularly, short regions (known as hot spots) with specific physicochemical properties are believed to initiate the self-assembly process by nucleating the aggregation reaction[199]. Because the intrinsic sequence properties that determine aggregation propensity are known[198,199], computational methods based on these properties can be used to analyze the aggregation propensities of protein sequences[200]. In this study, we will design protein variants with different physicochemical properties and predict their aggregation propensities using a combination of computational tools. This strategy will allow us to identify the specific protein regions that are responsible for αsyn's propensity to aggregate. Finally, we will design αsyn variants with specific predicted aggregation propensities, which can be experimentally tested to validate these predictions.

2.2.1. Hypothesis

Hypothesis: *specific regions in the gene encoding for alpha synuclein alter the protein's aggregation propensity.*

2.2.2. Methods

We collected protein sequence and functional information for αsyn (gene name: SNCA, UniProt ID: SYUA_HUMAN) and reported variants from different sources (i.e. databases, experimental information, computational predictor tools) as described below. These protein features were mapped into the protein sequence of αsyn. This information was used to explore the αsyn sequence and identify regions that potentially modulate the protein's aggregation propensity. Finally, a set of candidate mutations with the most dramatic affects on the protein's aggregation propensity was proposed.

2.2.2.1. Data collection

Natural αsyn variants (e.g. isoforms and pathogenic mutations), secondary structure elements, repeats and main mutagenesis experimental information were collected from the UniProt[24] database. Other experimentally validated functional features of αsyn, such as linear motifs, phosphorylation sites, and protein domains were extracted from the Eukaryotic Linear Motif resource for Functional Sites in Proteins (ELM)[106] database. Additional information on the features of αsyn was obtained from the AMYPdb[201] database, dedicated to amyloid precursor protein families and to their amino acid sequence signatures. αsyn protein sequence was used as query to search in the AMYPdb for homolog proteins and to search specific patterns in αsyn sequence that present homology with other amyloidogenic proteins. In addition, the AMYPdb was queried to verify if amyloid patterns from

other proteins matched αsyn sequence. All this information was complemented by an extensive literature search of mutagenesis experiments reported to alter aggregation propensity[196,202,203,204,205] and the aggregation propensity of αsyn isoforms[206,207] .

Functionally relevant regions of αsyn were predicted using sequence-based computational tools. Intrinsically disordered regions were predicted using IuPred[208], while disordered binding regions were predicted using ANCHOR[107] (Section 1.2.4.1). αsyn amyloid-forming regions were also predicted, because misfolding-prone proteins often self-assemble into both aggregates and amyloidogenic structures[209]. Amyloid prediction was performed using Waltz[210]. This tool uses an experimentally derived position-specific scoring matrix (combining sequence information, physicochemical and structural information) to assign residues a score that describes the propensity to mediate self-assembly and amyloid formation. Waltz's scoring matrix was derived from the biophysical and structural analysis of the amyloid properties of a large set of hexapeptides, and allows distinguishing between amyloid and amorphous aggregating sequences[210]. Three aggregation prediction methods were used to identify the regions of αsyn sequence that present aggregation propensity. The first method used is Aggrescan[211], which is based on an experimentally defined amino acid specific aggregation-propensity scale. Aggrescan's scale is derived from intrinsic sequence properties such as hydrophobicity, charge, packing density, and secondary structure propensity. This method identifies candidate aggregation-prone segments in

proteins (i.e. hot spots). We also used Zyggregator, which is based on similar sequence properties: hydrophobicity, charge, and the propensity of residues to adopt α-helical or β-sheet structure[212]. Zyggregator also incorporates in its predictions the environmental conditions in which the aggregation reaction is expected to occur. These conditions include pH, ionic strength, presence of denaturants, and polypeptide concentration in the solution; four factors that affect protein-aggregation rates[213,214,215]. Aggregation hot spots were also identified using TANGO[216]. This method is based on the physicochemical principles of secondary structure formation and it is based on the assumption that the aggregates' core regions are fully buried. The TANGO algorithm allows assessing the aggregation propensities of intrinsically disordered proteins because it is based not only on the proteins physico-chemical properties that determine aggregation such residue stretches having high hydrophobicity, high β-sheet propensity and low net charge[217] but also on the nature and frequency of aggregation-promoting nucleation stretches, which have been shown to be three times more frequent in globular proteins than for IDPs[216]. The fact that TANGO calculates the frequency of these nucleation stretches according to the type of protein (globular or intrinsically disordered) is important for accurately predicting aggregation propensities of αsyn, because it takes into account the fact that IDPs have a compositional bias that not only reduces secondary and tertiary structure, but that also reduces aggregation propensity[217].

2.2.2.2. Feature mapping

All protein sequence and functional information for αsyn and reported protein variants, mutagenesis experiments, along with all the important protein sites (e.g. phosphorylation, linear motifs) collected in the previous section were mapped into the protein sequence of αsyn. The purpose of this mapping was to integrate all information that might contribute to the identification of protein regions able to alter αsyn aggregation propensity.

2.2.2.3. Mutant design

Based on the information collected and mapped into αsyn's sequence as described above, point mutations in the gene encoding αsyn expected to modulate the predicted aggregation propensity were identified.

To identify candidate mutations decreasing the protein's aggregation propensity, we selected residues reported to have that effect. We include: i) charged residues, because they tend to increase protein solubility[218] (R preferred over D,E), ii) proline residues, because they function as aggregation breakers[211]. Conversely, we maintained positions of gatekeeper residues -typically glycines flanking aggregation hot spots- because they are considered to minimize aggregation[209,219,220].

The design principles used to identify the candidate mutations that increase the protein's aggregation propensity were based on mutagenizing aggregation-

breaker residues such as prolines and gatekeeper residues. Since extrinsic factors (such as pH decrease or a temperature increase) induce the formation of more ordered structures and accelerate αsyn fibrillation and aggregation[221,206,222], we assumed that point mutations of αsyn leading to a more ordered structure will also increase aggregation propensity. Thus, we also considered that replacing disorder-promoting residues (e.g. residues that favor more unstructured states of the protein) for order-promoting residues (e.g. bulky hydrophobic amino acids such as Ile, Leu and Val, aromatic amino acids such as Trp, Tyr, Phe, along with Cys and Asn[70,166]) may have an increasing effect on aggregation if such residues are located predicted disordered regions.

To test αsyn mutations we used the following method:

1. Point mutations were introduced in different regions of the protein sequence (e.g. disordered regions, regions predicted to aggregate, regions predicted to form amyloid, gatekeeper residues).

2. Aggregation propensity, amyloid propensity, intrinsic disorder and disorder binding regions of the protein variants were calculated.

3. αsyn variants showing at least a 50% decrease on the predicted aggregation propensity with respect to wild type αsyn were selected. Similarly, we selected αsyn variants showing the higher predicted aggregation propensity (up to 187%) than wild type αsyn .

4. Protein regions that affect αsyn aggregation propensity were defined based on the αsyn variants selected.

This method was systematically applied to predict the aggregation propensity of protein variants resulting from 436 point mutations in αsyn-encoding gene.

2.2.3. Results

The most important functional and structural features of αsyn were mapped on the protein's sequence (Figure 2-1). αsyn primary sequence can be divided into three regions:

i) *Residues 1-60*, containing four imperfect repeats that code for amphiphatic helixes (Figure 2-1A, cyan segments marked as H).

ii) *Residues 61-95*, containing the NAC (Non-Abeta Component) region and two additional protein repeats. The NAC region is a 35-residue segment rich in hydrophobic amino acids and seems to be involved in fibril formation[223]. .

iii) *Residues 96-140* - a predominantly charged and disordered region.

The first two regions are involved in binding to lipid structures, while the third region is involved in protein-protein interactions[2].

αsyn's protein coding exons are also shown in Figure 2-1A (numbered 2 to 6, top of the sequence). Exon 3 (residues 41-54) contains a region that prevents aggregation and it is not present in αsyn isoform 2-5[224]. Interestingly, the mutations

E46K and A53T, which are associated with hereditary forms of Parkinson's disease, are located in this exon. Additionally, αsyn isoform 2-4 does not contain exon 5 (residues 103-130), which increases aggregation[224]. We suggest that this increase in aggregation is due to the fact that that exon 5 is located in one of the predicted disordered regions, and thus decreases the protein's disorder content.

Aberrant phosphorylation of αsyn in association with disease development was reported[2]. Interestingly, the predicted disordered binding regions of αsyn (Figure 2-1B), are enriched in phosphorylation sites[106], which is consistent with previous works correlating phosphorylation sites with disordered regions of proteins[90]. Most of the mutagenesis experiments reported to alter αsyn's aggregation propensity[225], including previously characterized mutations found in hereditary cases of Parkinson's disease[203,204,205] (Ala30Pro, Ala53Tyr, and Glu46Lys), are located in the NAC region, where phosphorylation and ubiquitination are also rare (Figure 2-1B).

The predicted disordered regions are shown in Figure 2-1C. Three disordered regions were predicted by IuPred (marked in pink in the protein sequence). Two disordered binding regions were predicted by ANCHOR. The regions predicted to aggregate according to Aggrescan, TANGO and Zyggregator showing general consensus on their location are marked as red, green and cyan segments, respectively. Not surprisingly, the only amylodogenic region, predicted by Waltz (purple segment), matches well the fibril-formation NAC region (Figure 2-1A). Glycine residues in αsyn sequence, some of which act as gatekeepers of the

aggregation hot spots are also shown (Figure 2-1C). Interestingly, there are no proline residues (known aggregation breakers) in the first ~100 residues of αsyn's protein sequence. This ~100-residue segment spans virtually all the aggregation regions predicted by the different methods. The only predicted amyloid region in αsyn coincides with a predicted aggregation hot spot. αsyn's known motifs reported in the ELM database[106] are marked as dark gray regions in the protein sequence (Figure 2-1C). These motifs are believed to prevent the protein's aggregation propensity because they have a significant role in keeping native unfolded status of αsyn[226]. Querying of the AMYPdb database revealed that αsyn has 168 patterns shared with other aggregation-prone proteins from other families, such as the Islet Amyloid Polypeptide (IAPP) protein, the beta2 microglobulin, and cystatin C. In addition, there are 21 homolog proteins of αsyn reported in the AMYPdb.

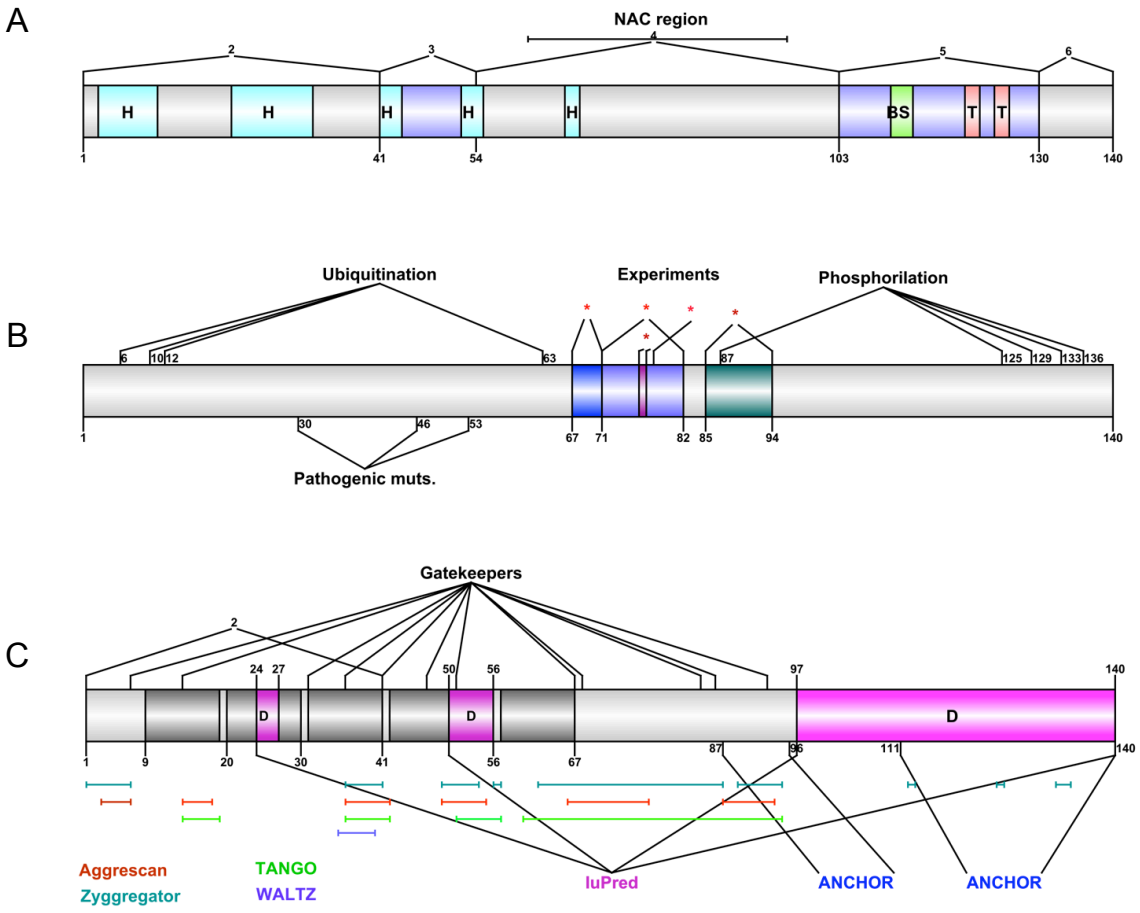

Figure 2-1. Representation of αsyn's main functional and structural features (obtained as described in Methods) and mapped into the sequence. A) Secondary structure elements (H=helix, BS=Beta Strand, T=TURN); highly amyloidogenic NAC region, and exon composition (numbered 2 to 6, alternating gray and purple; exon 3 is missing in isoform 2-5, while exon 5 is missing in isoform 2-4) B) Phosphorylation and ubiquitination sites; pathogenic mutations, and other relevant rationally-designed mutations. C) Disordered regions (pink, predicted by IuPred); disordered binding regions (predicted by ANCHOR); amylodogenic region (predicted by Waltz); aggregation-prone regions (predicted by Aggregscan, Zyggregator, and TANGO) and positions of glycine residues, potential gatekeepers of aggregating regions. Dark gray regions correspond to reported protein motifs (from ELM database).

2.2.3.1. Mutations in specific regions of αsyn alter its aggregation propensity

The aggregation propensity (AP) scores predicted by TANGO for αsyn wild type (AP score = 896.1, dotted line) and for the 436 variants containing single-residue substitutions (stars) are shown in Figure 2-2 (top). Four "aggregation dips" can be identified by the local minima of AP scores corresponding to specific protein regions. Point mutations in these regions tend to have more dramatic effect on the AP scores than those outside of the aggregation dips, resulting in AP scores significantly lower than wild type αsyn. Moreover, the nature of the mutation has little effect the protein's AP as residues with different chemical properties result in similar AP scores. In addition, protein variants resulting from mutations disrupting known αsyn motifs[106] (dark segments in Figure 2-1A) also seem to have marginal effect on the protein's AP. This observation is not consistent with previously reported evidence showing that motifs may have significant role in maintaining αsyn in natively folded state[226].

The representation of αsyn's sequence with the regions predicted to be aggregation prone, amylodogenic, and disordered is shown in Figure 2-2 (bottom). Interestingly, the aggregation regions predicted using different methods (Aggregscan, TANGO and Zyggregator; shown as thin colored segments, Figure 2-2) match the aggregation dips resulting from point mutations located in those regions. Four *consensus* hot spots spanning the protein sequence (corresponding to a consensus between the aggregation predicted regions and the local minima of AP

scores) were predicted: *hs1* (residues 14-19), *hs2* (residues 34-43), *hs3* (residues 49-57) and *hs4* (residues 66-79). The predicted disordered regions (in grey) and the predicted disordered binding regions (in blue) appear to flank the hot spots, suggesting that aggregation prone regions are not compatible with structural disorder. Consistent with previous observations[209], small, hydrophobic residues such as Val, Ala, and Tyr tend to increase the protein's aggregation propensity regardless of the position of the mutation with respect to sequence elements such as motifs, repeats and disordered regions. Prolines also function as aggregation breakers when located in the protein's hot spots. In summary, the protein AP is affected by mutations in these aggregation hot spots and the aggregation propensities of αsyn variants resulting from point mutations inside aggregation hot spots are similar in value. Thus, in order to decrease the protein's AP, it is important to mutate residues inside the consensus hot spots, while also considering that the closer the position of the mutation to the hot spots' ends, the lower the effect of the mutation on the AP. Our results also confirm that mutations of gatekeeper glycines (those flanking the hot spots' ends) tend to increase aggregation. The use of gatekeeper residues seems to have evolved as a strategy to minimize aggregation[209,219,220]. Furthermore, gatekeepers are also thought to determine chaperone selectivity for highly aggregation-prone protein regions.

The analysis on the different αsyn protein regions lead to propose specific mutations of the protein, which have the most dramatic effects on the aggregation propensities according to predictions obtained using TANGO. The Lys23Val

substitution was chosen as control because it does not belong to any predicted aggregation hot spot, or to any predicted disordered region, thus we assume should not affect the disorder content of the protein). In fact, this mutation results in an AP score (AP score = 892.7) very similar to that of wild type αsyn.

Mutations in the protein regions defined by hot spots *hs4* and *hs2* result in the lowest aggregation propensity scores. To analyze *hs4*, we propose to mutagenize Val71 to charged residues such as Lys and Arg (AP score = 408.8 for both). Mutagenizing Val74 to Asp will result in a similar aggregation propensity score (AP score = 406.5). To investigate *hs2*, we propose to test the mutation Val37Pro, which was predicted to result in the lowest aggregation propensity (AP score =610.01). In addition, combining pairs of proposed mutations to generate the double mutant Val37Pro/Val71Lys αsyn was observed to have a synergistic effect on aggregation (AP score = 121.7). A similar result was obtained with the double mutant Val71Arg/Val37Gln αsyn (AP score = 123.8).

The highest aggregation propensity score was observed when mutations were introduced at position 50 to substitute histidine (a charged residue that should oppose aggregation) with the hydrophobic residues Val or Ile (AP score =1680.7 and 1678.1, respectively). Interestingly, this position marks the beginning of a predicted disordered region, and considering that Val and Ile are order-promoting amino acids[70,166], our results are in agreement with previously reported evidence suggesting that an increase in protein structure enhances αsyn's aggregation[221,206,222]. The order-promoting Tyr54Val substitution was also

observed to cause a dramatic increase in AP (AP score =1205.1). The combined effect of double mutants with opposing AP tendencies results in a decrease in AP (Val71Lys/His50Val, AP score = 1193.7). Finally, because mutating glycines acting as potential gatekeepers of aggregation hot spots also increases the AP, we propose mutating glycine at position 14, which flank the *hs1*, into a Val (AP score =1361.9) or into an Arg (AP score = 1026.4).

In summary, we presented an effective strategy to identify αsyn regions that alter the protein's AP based on literature and experimental information and using different protein sequence analysis tools. Interestingly, selecting mutations belonging to the predicted aggregation hot spots resulted in more dramatic effects on the AP than selecting mutations that disrupt αsyn's motifs and repeats, which were previously reported to increase aggregation[226]. In addition, our results suggest that point mutations into "order-promoting" residues[70,166] (i.e. residues that decrease the protein's disorder content) enhance the protein aggregation propensity. This approach allowed to rationally designing αsyn variants with specific APs (Table 2-1). These variants can be used to experimentally investigate αsyn aggregation propensity and to define its sequence-structure relationship in the context of protein aggregation. In fact, experimental analyses of four of the proposed single-residue variants αsyn (Val71Lys, Val37Pro, Tyr54Val, His50Val) demonstrated having solubility values in mammalian cells that confirm aggregation predictions obtained from this study (data not shown, LS and NP personal communications).

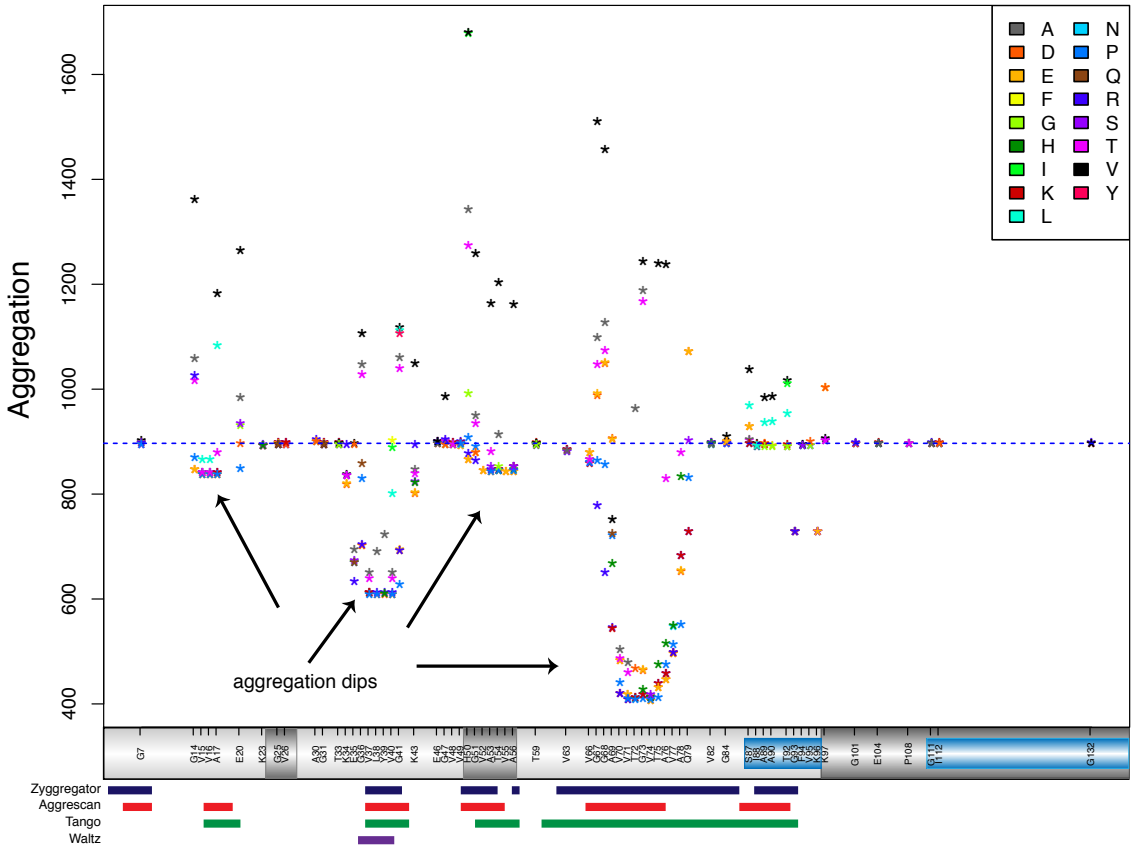

Figure 2-2. Aggregation propensity scores of αsyn (wild type) and of αsyn containing single-residue susbtitutions. Top: The aggregation score was calculated with TANGO, the dotted line represents the aggregation propensity obtained for αsyn wild type. Each star corresponds to the aggregation score (y-axis) of a αsyn variant containing a single amino acid substition (x-axis). Bottom: aggregation regions of wild type αsyn. Dark grey segments correspond to predicted disordered regions (IuPred) and blue segments correspond to predicted disordered binding regions (ANCHOR). Aggregation regions predicted using Zyggregator (blue), Aggrescan (red), Tango (green) and Waltz (purple) are also reported.

Table 2-1. Aggreagation propensity of αsyn variants

Mutation (single or double)	Predicted aggregation propensity (AP) effect
Lys23Val	Control
Val37Pro; Val37Gln; Val71Lys; Val71Arg; Val74Asp	Decrease AP
Val37Gln/Val71Arg; Val37Pro/Val71Lys	Decrease AP
Gly14Val; Gly14Arg; His50Val; His50Ile; Tyr54Val	Increase AP
His50Val/Val71Lys	Increase AP

AP = Aggregation propensity

2.3. Classification and functional specificity of the Racemase protein family

Racemases are enzymes that convert L-amino acids into D-amino acids by changing the stereochemistry of the chiral alpha-carbon atom[5]. Until recently, the roles of D-amino acids in bacterial physiology were limited to bacterial spore germination, regulation of the peptidoglycan (PG) cell wall and cell growth[227,228,229]. However, recent findings have demonstrated that noncanonical D-amino acids (NCDAAs) are synthesized and released to the environment by bacteria from diverse phyla. It was suggested that NCDAAs function as signaling molecules to mediate bacterial communication in extreme environments, such as under conditions of nutrient deprivation[230]. Interestingly, NCDAAs can communicate with cells that synthesize NCDAAs as well as with neighboring cells, and with cells from different bacterial species[6]. This rapid diffusion of NCDAAs enables a quick and

synchronized response from the whole bacterial population. The molecular mechanisms that regulate production of NCDAAs remain unclear. Moreover, racemases are poorly characterized and generally considered substrate-specific. Lam *et al*., however, suggested that some racemases can have multisusbtrate specificity[6]. They reported a *Vibrio cholerae* racemase that accounts for the accumulation of D-Met, D-Leu, D-Ile, and D-Val, and racemizes 10 different D-amino acids. This Broad Spectrum Racemase (BsrV), however, was annotated in the UniProt database[231] as Alanine Racemase 2 (ALR2_VIBCH), because it was thought to be only capable of racemizing alanine. *V*. *cholerae* produces another alanine-specific racemase (ALR1_VIBCH). Thus, the current UniProt annotation of these two proteins does not distinguish between the type of D-amino acid they produce: they are both allocated Enzyme Commission Number (EC=5.1.1.1) indicating that they produce D-Ala. These two racemases share similarity in protein sequence and structure, and both use pyridoxal 5'-phosphate (PLP) as cofactor to catalyze the racemization reaction, yet they have been experimentally reported to exhibit different substrate specificity[6]. Horcajo *et al.* suggested that, in fact, many bacteria may have mis-annotated racemases, which have broader substrate specificity than what originally thought. This mis-annotation is likely to be due to the high sequence and structure similarity of racemases and to the use of PLP as cofactor, which is a common feature in multiple bacterial proteins.

We proposed to identify the differences in the protein sequences of *V. cholerae* ALR1 and BrsV that could account for their different substrate specificity.

The resulting molecular footprint could be used to design protein variants of BrsV with different substrate specificities for NCDAA production that could aid exploring their role in bacterial physiology. NCDAA production has also had increasing application in the pharmaceutical industry, biotechnology, immunodiagnostics, and food industry[232]. This molecular footprint could also be used to analyze newly discovered proteins[1], thus allowing the proper classification of racemases.

2.3.1. Hypothesis

Hypothesis: *Specific protein residues in racemase sequences determine the protein's substrate specificity.*

2.3.2. Methods

We collected racemase protein sequences from different bacteria and aligned them. To identify residues putatively responsible for substrate specificity, we started by detecting positions showing differential conservation patterns for alanine racemases and for broad-spectrum racemases in the multiple alignment (MSA). These specificity determining positions, or SDPs (Section 1.1.2.2.) were identified using the Xdet[233] and S3Det[44] methods implemented by the JDet tool[45]. S3Det, concomitantly with the SDP prediction, provides an automatic splitting of the alignment into subfamilies[1]. This subfamily classification of racemase proteins was compared to the corresponding phylogenetic tree of the whole MSA. A molecular footprint of the residues putatively involved in the substrate specificity of the racemase protein family was defined based on those positions differentially

conserved in the MSA subfamilies. This footprint of residues was also used to identify 77 putative multi-substrate specific racemases from different organisms that had been probably miss-annotated previously as being specific racemases.

2.3.2.1. Dataset

The dataset was constructed by joining the results of two protein BLAST[234] searches selecting ALR1 and BrsV from *V. cholerae* as query sequences, and performing non-redundant searches with a default parameters and a threshold of 1E-10. These independent searches resulted in highly overlapping protein sets (3,967 proteins for ALR1, and 3,595 proteins for BrsV), stressing the high degree of homology of alanine racemases. Additionally, we included in the MSA protein sequences for which experimental or structural information regarding substrate specificity was available (F. Cava, personal communication), including alanine racemases from *E. coli*, *B. subtilis, A. hydrophila*, and *A. baumanii*. The resulting sequences were filtered selecting only those mapped into the UniProt database, yielding a preliminary set of 1,355 protein sequences.

2.3.2.2. Multiple Sequence Alignment

The protein dataset was aligned using multiple sequence alignment tool MUSCLE[235] (v3.8.31). The resulting MSA was filtered out for empty columns and for redundant proteins (using a 95% redundancy threshold) using the Jalview[236] tool (v2.7), resulting in 137 sequences in the final protein dataset.

2.3.2.3. Identification of Specificity Determinant Positions (SDPs) and protein subfamilies

Identification of SDPs was carried out using Xdet and S3Det implemented in JDet. JDet allows extracting, visualizing and manipulating fully conserved positions and family dependent positions in MSAs[45]. Xdet and S3Det were applied to explore the residue conservation pattern of the protein dataset. Xdet compares the mutational behavior of a given position in the MSA to the mutational behavior of the whole alignment[233]. A matrix containing physicochemical similarities represents the mutational behavior for all pairs of amino acids at a given position. Similarly, a matrix containing the overall similarities for all pairs of proteins encodes the mutational behavior of the whole alignment. The comparison of these matrices provides a score for the position of the MSA, where highest scores are selected as predicted SDPs. S3Det is based on a vectorial representation of the MSA in a high dimensional space followed by a Multiple Correspondence Analysis treatment for dimensionality reduction. Each protein is presented as a vector, and vectors representing proteins with high sequence similarity are clustered in the same regions of the sequence space, allowing for the identification of the internal organization in subfamilies or subgroups[1]. A similar vectorial transformation for the individual MSA positions results in a residue space where SDPs are located in the same regions were the clusters representing their associated subfamilies are located. Thus, S3Det not only detects SDPs but also defines the subfamily composition of the MSA according to positions having differential conservation

patterns within the MSA[44]. Such positions may be conserved in a given subgroup but not in another, or the conserved amino acid might be different among the subgroups[1].

A phylogenetic tree was constructed from the MSA protein sequences to compare it with the partition of the MSA into subfamilies generated by S3Det. The tree was based on the neighbor joining method using substitution matrix BLOSUM62 and was built using the Jalview[236] tool (v2.7). The manipulation and display of the resulting phylogenetic tree was performed using iTool[237].

2.3.3. Results

2.3.3.1. The racemase protein family can be subdivided according to substrate specificity

Applying S3Det resulted in a subdivision of the MSA of the racemase proteins into three subfamilies. The protein space representing the internal organization in subfamilies (clusters) is shown in Figure 2-3. Subfamily 1 (in red) is composed of 84 proteins and includes ALR1 from *V. cholera*, ALR1 and ALR2 from *E. coli*, which have all been experimentally reported to be alanine specific racemases[6]. Subfamily 2 (in blue) comprises 13 proteins, including BrsV (ALR2 from *V. cholera*) and alanine racemases from *A. hydrophila* and *A. baumanii*, all reported to use various amino acids as substrates (In preparation). Subfamily 3 (in green) included the remaining protein sequences in the dataset. There were 15 outlier proteins that could not be assigned to any of the subfamilies. We assumed that proteins in subfamily 1

corresponded to substrate specific alanine racemases, while proteins in subfamily 2 have a broader substrate spectrum. Subfamily 3 did not include proteins for which experimental information on their substrate was available; therefore it was excluded from the analysis.

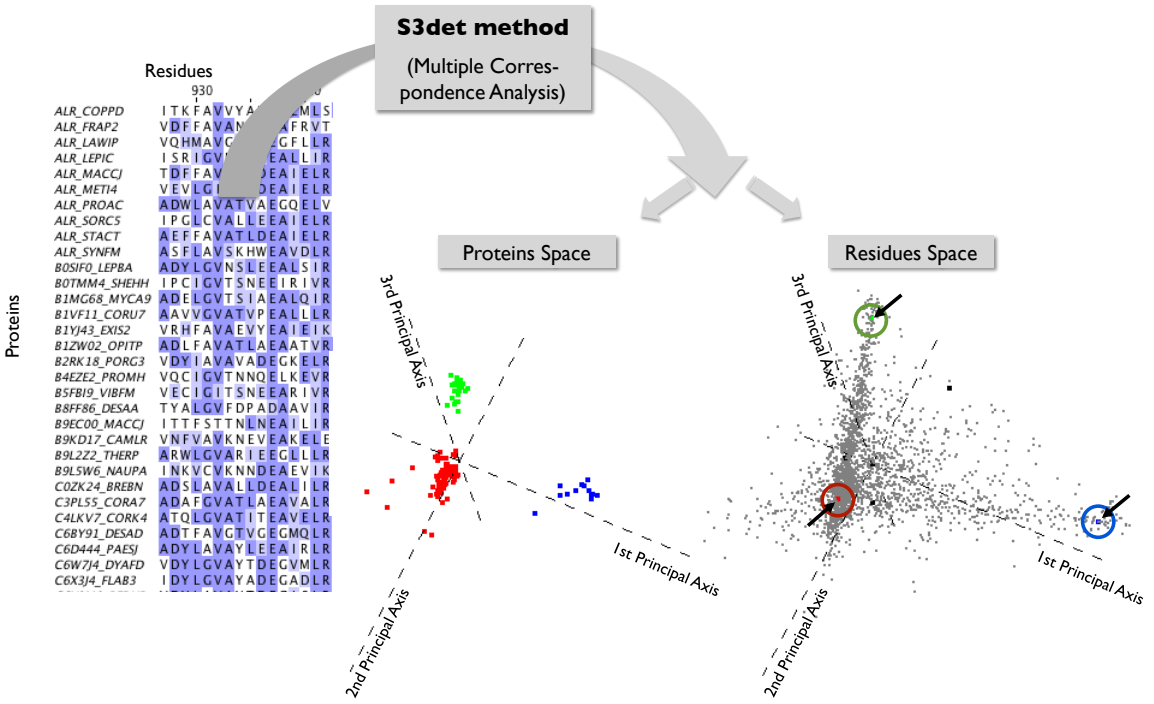

Figure 2-3. Schematic representation of S3Det results visualized with JDet. S3Det is applied to the multiple sequence alignment of the racemase family (a fragment of it is shown). The three-dimensional projections of the reduced protein and the residue spaces are shown. The protein space represents similar proteins clustered in the same spatial regions assumed to correspond to the different subfamilies (marked in red, blue and green). SDPs are located in the corresponding regions of the residue space where the clusters of subfamlies are located in the protein space. The centers of mass of each protein subfamily are represented by circled dots.

As expected, the subfamily classification implicit in the phylogenetic tree (Figure 2-4) is in high correspondence with the subfamily classification provided by S3Det.

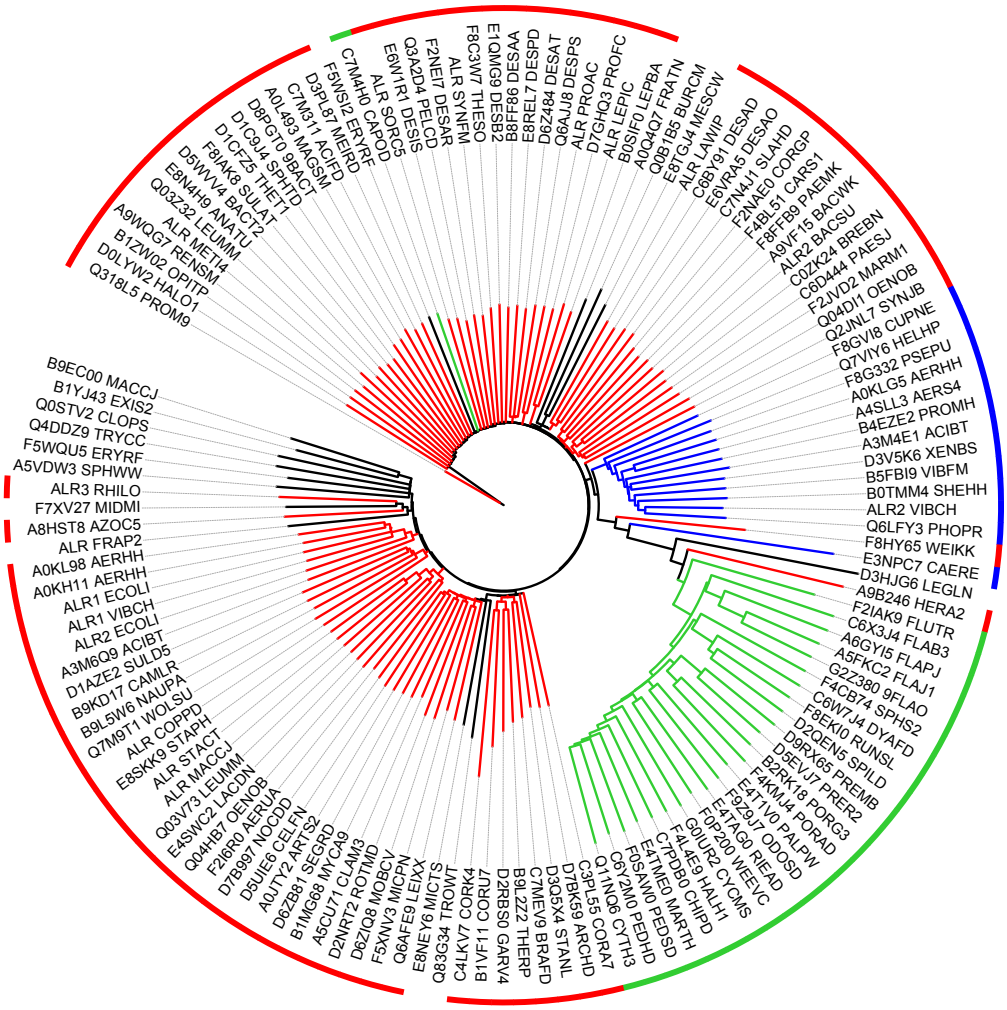

Figure 2-4. Phylogenetic tree of the multiple sequence alignmen (MSA) of the racemase familiy using neighbour joining with substitution matrix BLOSUM62. The three subfamilies defined by S3Det are shown: subfamily 1 (blue), subfamily 2 (red) and subfamily 3 (green). Outlier protein sequences from the MSA are shown in black. Figure generated with iTol[237].

2.3.3.2. Predicted specificity determining positions in the racemase protein family

The conservation-based predicted functional sites identified by the Xdet and S3Det methods resulted in 31 consensus SDP candidate positions. Each SDP was manually inspected, and only those SDPs having a clear differentiated residue pattern between the two subfamilies of interest were selected. SDPs are considered as "important" for the protein's specificity, thus, they are likely involved in protein interactions, involved ligand binding, or part of the catalytic site. However, nothing can be inferred on the functional role of these positions. Therefore, SDPs were also examined from their structural and functional context to distinguish only those SDPs potentially involved in substrate specificity. This information was based both on literature and expert knowledge. Alanine racemases are formed by a head-to-tail association of two monomers, where each monomer is composed of an N-terminal α/β barrel domain and an extended β-strand domain at the C-terminus[238] (Figure 2-6). The active site in each monomer is located at the center of the α/β barrel and contains the PLP co-factor covalently connected to a lysine. The catalytic mechanism involves the same lysine and a tyrosine contributed by the opposite monomer[239,240,241]. Residues involved in the entryway of the active site and the PLP binding site are located in the loops of the α/β barrel domain of one monomer and residues from the C-terminal domain of the other monomer[242]. Interestingly, all the structurally and functionally relevant reported positions were included in the 31 candidate SDPs. This mapping of the structural and functional information into the

SDPs candidate positions, together with the previous manual assessment of the amino acid patterns, resulted in a final molecular footprint of 16 consensus functional sites putatively related to the substrate specificity of racemases (Figure 2-5).

Figure 2-5. Substrate specificity determining positions (SDPs) in the racemase family. In the top, the 16 differentially conserved positions in subfamily 1 (alanine specific). In the bottom, the 16 differentially conserved positions in subfamily 2 (broad-spectrum specificity).

2.3.3.3. Racemase's subfamily specificity can be achieved by selective mutagenesis

The molecular footprint classifying racemases according to substrate specificity allowed the identification 16 putative function-specific residues related to substrate specificity. Accordingly, 16 point mutation variants of the BsrV gene were engineered to test their substrate binding: Cys70Ala, Arg119Ala, Arg121Ala,

Ala165Lys, Asn167Ala, Gly169Ala, Asn174Ala, Pro206Asn, Tyr208Ala, Lys216Ala, Tyr264Ala, Gly263Ile and Asn38Ala. Preliminary results on substrate binding assays of these BsrV variants report that certain point mutations result in complete loss of substrate binding. Other point mutations, however, result in BsrV variants that selectively limit the binding of the different substrates (i.e. the L-amino acids) (In preparation). Thus, it is possible to modulate racemases' binding to the different substrates by selective mutagenesis. The 16 residues putatively shared among broad-spectrum racemases mapped into the BsrV homodimer structure are shown in Figure 2-6.

2.3.3.4. Detection of putative broad-spectrum racemases in bacterial genomes

S3Det was also applied to the MSA of the pre-filtered dataset of 2,967 proteins reported in Section 2.3.2.1. The resulting subfamily division of this enlarged set was compared to subfamily classification previously obtained for the filtered set. This allowed for the classification of the new proteins into the previous "broad-spectrum" or "specific" categories. The assignment of the new proteins was obtained by assessing the enrichment of the new subfamilies in "broad-spectrum" or "specific" racemases as inferred from the original classification. This strategy allowed the identification of 77 BsrV-like racemases from different organisms. This set of putatively broad-spectrum racemases is being currently characterized; and two of them have been crystalized and their broad spectrum experimentally demonstrated (F. Cava, personal communication, In preparation).

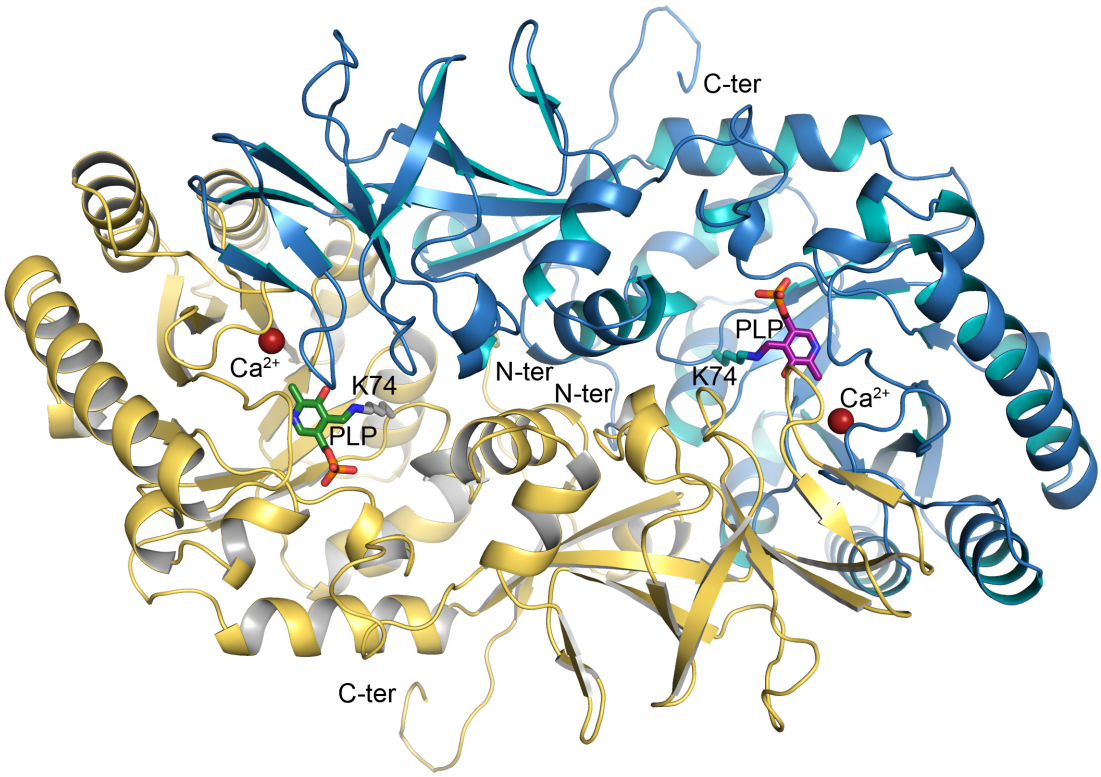

Figure 2-6. Mapping of the 16 substrate specific residues from the moleuclar footprint into the BsrV homodimer.

2.4. Discussion

The two case studies presented show different applications of computational tools to obtain functional information from protein sequences and highlight the advantages they offer to guide experimental assays.

In the αsyn protein, there is an urgent need to understand the molecular mechanisms underlying αsyn aggregation in living cells[195]. Recent developments have ensured ways to quantify αsyn solubility in living cells, thus, a necessary advancement to this quantification implies testing αsyn variants that efficiently modulate aggregation. The approach presented allowed us to systematically explore the effects of 354 single residue mutations in αsyn on its aggregation properties to enable the selection of the best candidates to be further tested in experimental settings in a time and resource efficient manner. We believe this is an adequate strategy that could help identifying regions modulating αsyn's aggregation in living cells. The identification of the mechanisms that govern amyloidogenicity, aggregation and toxicity of αsyn will likely contribute to the development of therapeutic strategies to prevent and treat neurodegenerative diseases. Moreover, the strategy implemented is not exclusive for studying aggregation of pathologically related proteins such as αsyn; this approach can be adopted for exploring aggregation of proteins in other contexts too. Protein aggregation is a major issue during biotechnical production and purification of proteins or engineered polypeptides used as drugs[243]. Additionally, aggregation of proteins in therapeutic formulations has also been reported to reduce drug effectiveness and even to induce serious secondary effects[244]. Thus, there is imminent need to understand and control aggregation processes in cells.

Our approach was based on choosing protein residues whose biophysical properties are predicted to impact AP or that have been reported to affect AP. An

alternative strategy could involve designing candidate mutations according to other criteria, for example, by aligning all protein sequences of the αsyn family, and performing an analysis distinguishing those proteins that aggregate from those that do not. Then, perform an analysis of residue conservation to determine aggregation specificity residues to be used as candidate mutations. Additionally, experimental results could provide feedback for fine-tuning our approach.

In the study of the racemase family, we implemented a strategy for characterizing their substrate specificity. There has been a crescent need for understanding the role of NCDAAs as crucial players in diverse aspects of bacterial physiology[230,](F. Cava, personal communication) and for biotechnological purposes[232]. Thus, understanding the substrate specificity of racemases producing NCDAAs may help in the their characterization, especially since the current annotation of this protein family does not distinguish between racemases acting on a specific substrate from those having multisubstrate specificity.

To characterize substrate specificity, we used methods that identify evolutionary conserved positions in proteins and integrated them with functional and structural information available for a few experimentally characterized racemase proteins. The presence of SDPs in protein regions related to functional and interaction specificity was recently shown to be a widespread phenomenon [44]. Moreover, the role of SDPs in controlling the functional specificity of in protein families has been experimentally demonstrated by mutating the corresponding residues[41,245].

Our strategy allowed the identification of a residue molecular footprint distinguishing alanine-specific racemases from those racemazing more than one amino acid. In addition, this set of residues was used to propose 16 point mutation variants of BrsV that were engineered to experimentally test their substrate binding. Several of these mutants proved to have altered substrate binding, paving the way to designing racemases “a la carte”, depending on the choice of D-amino acid wanted to produce. This design is currently being refined in the laboratory with the aim of producing D-amino acids according to specific biotechnological needs and its implementation is less expensive than current techniques (F.Cava, personal communication).

Additionally, 77 BsrV-like racemases from different organisms were identified using the molecular footprint of putatively specificity determining positions. Thus, our approach has proven to be effective not only for the characterization of the racemase protein family, but also for the potential production of NDAAs with specific biotechnological ends.

Chapter 3

Intrinsic disorder at genomic scale: Genome-wide analysis of intrinsic disorder and its implication in specific protein functional classes in *Arabidopsis thaliana*

3.1. Introduction

As part of the strategy to identify the functional role of intrinsic disorder in proteins from different biological systems, we proposed to perform a genome-wide analysis of intrinsic disorder and its relation to function in *Arabidopsis thaliana*.

As described in Section 1.2, the relationship between organismal complexity and intrinsic disorder remains unclear. It is currently accepted that prokaryotic and

eukaryotic organisms have different levels of disorder (in agreement with their organismal complexities)[246]. However, a number of studies also showed that, both prokaryotes and eukaryotes, when faced with adverse conditions in their environments use proteins enriched in intrinsic disorder to communicate and interact (with other cell types in the first case, and with the environment in the second case)[124].

Based on these premises, we hypothesized that *A. thaliana* could rely on protein disorder to respond to changes in environmental conditions such as cold and drought. Because plants are sessile organisms, they cannot escape from threatening conditions as other organisms do. As a result, phenotypic plasticity (i.e. the capacity to adapt their phenotype to changing conditions) is particularly important in plants to adapt and survive in rapidly changing environments. Phenotypic plasticity requires the integration of external information with the basal genetic and developmental programs, and it is achieved in plants through complex signaling networks [20]. Thus, we propose that plants use disorder as a simple and fast mechanism, independent of transcriptional control, for introducing versatility in the interaction networks underlying these biological processes to quickly adapt and respond to challenging environmental conditions.

While intrinsic disorder in the human genome has been vastly characterized[72,247,248], the current knowledge of the implications of disorder in plants is limited to a few case studies belonging to very specific protein families (namely LEA proteins and dehydrins) and confined to very specific biological

functions[249,250,251]. Even if some of these works demonstrated that proteins involved in signaling and environmental adaptation contain disordered regions (a notion that may imply a relation between protein disorder and phenotypic plasticity)[252,253,250], a whole-proteome analysis of the functional role of protein disorder in plants that could support this hypothesis has never been conducted.

To address this question, we assessed the level of intrinsic disorder (IDPs and IDP regions) in *Arabidopsis thaliana*, the most widely used model organism in plant biology, and focused on the biological functions that IDPs and IDP regions perform in this organism. In addition, we also compared the disorder content of functional classes shared between *A. thaliana* and human proteins. Our large-scale comparative analysis of protein disorder in these two organisms provided insights on the specific functional roles that this phenomenon plays in *A. thaliana*.

3.2. Hypothesis

Intrinsic disorder provides a mechanism to increase Arabidopsis thaliana's ability to adapt to the environment.

3.3. Methods

To assess the intrinsic disorder level of *A. thaliana* and Human at organismal level, we compiled two datasets containing the protein sequences and their corresponding Gene Ontology (GO) annotations of *A. thaliana* and human

proteomes. We calculated the disordered regions of each protein sequence in the datasets using different disorder prediction methods as described below. We then performed an enrichment analysis of the functional annotations (Gene Ontology terms) for disordered proteins in *A. thaliana*. Finally, we carried a similar functional enrichment analysis to compare disordered proteins in *A. thaliana* and Human. A schematic view of the workflow is shown in Figure 3-1.

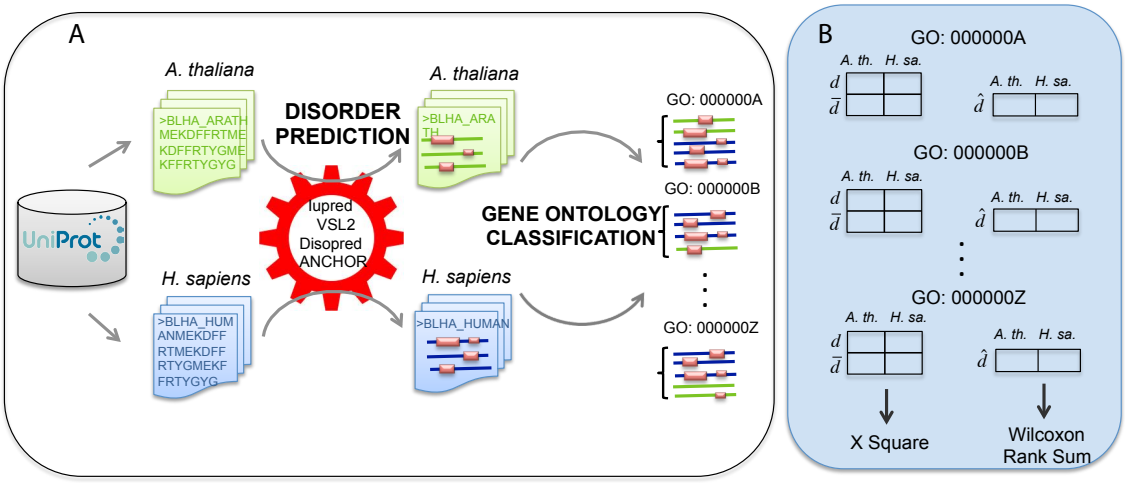

Figure 3-1. Schematic representation of the methodology used to study protein disorder in *A. thaliana* and its comparison with *H. sapiens*. A) For each organism (*A. thaliana* (green) and Human (blue)) protein sequences and their corresponding Gene Ontology annotations were retrieved from Uniprot. For each protein, disordered regions (pink) were calculated using 3 different methods (IuPred, VSL2 and Disopred), and disordered-binding regions (DBRs) were predicted using ANCHOR. Proteins were assigned to Gene Onotology (GO) functional classes. Functional classes significantly enriched in disordered proteins were identified for *A. thaliana*. B) Analysis of functional classes shared between *A. thaliana* and Human. For each GO functional class, a comparative analysis of the disorder levels of the proteins of each organism was performed using different measures for quantifying disorder. For the disorder measures assigning a binary classification of disorder of proteins, contingency tables were constructed to report the counts of disordered and not-disordered proteins in both organisms. The Chi-squared test was applied to evaluate the signicance of the differences in the reported counts. For the disorder measures quantifying disorder content of proteins in each GO class, the tables contain the average disorder content for each organism, and a Wilcoxon Rank Sum test was applied to measure the signifance of the differences of the mean disorder content.

3.3.1. Datasets

The datasets for the analysis were constructed extracting the proteome sequences of *A. thaliana* and *H. sapiens* from the Protein Knowledgebase[24] (UniProtKB, release 2011 04). We used the search engine of this resource to look for "A. thaliana" and "H. sapiens", and selected the "complete proteome" option, resulting in two sets of 32,764 and 35,346 sequences including both canonical proteins and isoforms. These datasets were filtered out for repeated, fragmented and proteins containing non-standard residues (such as Selenocysteine) and ambiguous residues (e.g. B, X, Z), which may not be tractable by certain disorder prediction algorithms.

The final sets contained 32,398 proteins for *A. thaliana* (from 31,304 genes) and 35,244 proteins for *H. sapiens* (from 20,154 genes), respectively.

3.3.2. Functional annotations

In order to assign functional terms to the protein sequences in the datasets, we adopted the functional vocabulary defined by the Gene Ontology Consortium[254]. (release 2011 04) Gene Ontology (GO) terms describe different functional aspects of gene products and are divided into three independent categories (subontologies): "BP: biological process", "CC: cellular component" and "MF: molecular function". The GO annotations for our sequences were also retrieved from UniprotKB. Terms that were labeled by the GO Consortium as "obsolete" were filtered out from the analysis. *A. thaliana* genes were annotated with a total of 4,278 GO functional terms from the

three subontologies and human genes were annotated with a total of 8,836 GO terms.

The controlled vocabulary defined by GO is specifically designed to be species-independent and to include only terms applicable to both prokaryotes and eukaryotes, single and multicellular organisms. Thus, with the proper handling of the common GO terms, this functional classification enables the comparison of the underlying molecular biology of gene products coming from such distinct taxonomic groups as *A. thaliana* and *H. sapiens*[255,256].

The Gene Ontology is structured as a directed acyclic graph where the terms are related by parenthood relationships. It can be thus navigated from very general (e.g. "enzyme") to more specific functions (e.g. "coenzyme F390-G hydrolase activity"). Generally speaking, the original GO annotations contain only the most specific terms that can be assigned to a given protein. In this analysis, we expanded the original set of GO terms annotated for a given protein by including all the ancestors of these GO terms. Thus, we ensured that any pair of proteins could be functionally compared at the GO level where they share an annotation. In other words, a given protein annotated as "enzyme" and another protein annotated as "coenzyme F390-G hydrolase activity" could be compared at the level of their common term "enzyme".

This term expansion of our annotations resulted in 6.410 GO terms (of all three subontologies) for *A. thaliana* and 12.690 GO terms for *H. sapiens*. From all

these terms, 4.380 annotated both *A. thaliana* and *H. sapiens* proteins, and hence only those were used for the comparative analysis.

3.3.3. Protein disorder prediction

The workflow for the prediction of intrinsic protein disorder was implemented with ad-hoc scripts (developed in Perl programming language) which ran three different tools: Disopred[171] v2.4, VSL2[257] and IuPred[208]. The first two disorder predictors are based on linear support vector machines. The latter is based on the pairwise energy content estimated from residue composition. Section 1.2.4.1 includes a more detailed description of these prediction methods. These methods take a single protein sequence as input and provide as output a disorder probability in the 0.0 – 1.0 range for each residue. In order to convert these values into a binary ("ordered" vs. "disordered") prediction at the residue level, we used the default threshold for each predictor (0.5 for VSL2, and IuPred and 0.05 for Disopred).

For each protein in the two datasets (*A. thaliana* and *H. sapiens*), we defined two disorder metrics: i) relative disorder content (i.e. the percentage of disordered residues in whole protein), and ii) number of long disordered regions (LDR, which are defined as number of protein regions with at least 30 consecutive disordered residues). These two metrics represent two different disorder criteria and are typically used in the disorder field[171,258].

Additionally, we extracted the disordered regions predicted to undergo disorder-to-order transition upon binding. We developed a script based on the

ANCHOR method[107], described more in detail in Section 1.2.4.1. This method identifies potential sites of molecular attachment, hence it can be used to predict regions involved in protein-protein interactions located in disordered regions of the protein sequence[107]. Similar to the disorder prediction methods, using the protein sequence as input, it identifies the disordered binding segments.

3.3.4. Evaluating the disorder in Gene Ontology functional classes

This analysis can be divided into two different parts: i) identify the GO classes significantly enriched in disordered proteins in *A. thaliana*, and ii) identify GO classes differentially enriched in disordered proteins for *A. thaliana* with respect to Human.

To evaluate whether a given GO class was significantly associated to disordered proteins in *A. thaliana* we quantified disorder of all classes. We then conducted an "enrichment analysis" test[259] as implemented in the Database for Annotation, Visualization and Integrated Discovery[260] tool (DAVID, v6.7). DAVID is widely used in the scientific community for the systematic and integrative analysis of gene lists, and allows identifying enriched biological annotations, with particular emphasis on GO terms. The disorder of each given GO class was quantified by counting the number of proteins with at least one long disordered region (LDR) according to the different disorder prediction methods (IuPred, Disopred and VSL2). The input parameters used for DAVID were: *Background:* "Arabidopsis thaliana". *Gene Ontology subontologies:* "GOTERM BP ALL", "GOTERM MF ALL", and "GOTERM

CC ALL". A "Functional Annotation Chart" was generated from this enrichment analysis listing all the GO annotation terms and their corresponding associated genes. This list was filtered by p-value (using the correction by Benjamini, p-val ≤ 0.05) and by minimum number of genes belonging to each annotation term (count =2).

To perform the comparative analysis of the disorder of the GO classes common to *A. thaliana* and Human, we quantified the "disorder" of a given GO class in each organism with the same criterion described above. Then, a 2x2 contingency table was constructed containing the number of disordered proteins for each of the two organisms (A. th. and H. sa.) and the complementary counts (number of "non-disordered" proteins according to that criteria), as shown in Figure 3-1B. We measured the significance of the difference between the observed and the expected frequencies of disordered proteins in *A. thaliana* and *H. sapiens* with a Pearson's Chi-squared test with Yates' continuity correction[261]. We considered only the classes for which the number of disordered proteins in *A. thaliana* was higher (5% or more) than the "expected" value reported by the Chi-squared test. This method to filter the results allowed us to identify the functional classes for which the difference in disorder was positive for *A. thaliana*. With this procedure, we assigned a p-value to each GO functional class, which was corrected using the Benjamini & Hochberg multiple testing correction[262]. Consequently, GO classes with low p-values correspond to those significantly enriched in disordered proteins in *A. thaliana*

when compared with Human. All statistical analyses to estimate significance were implemented in the statistical analysis programming language R[263].

There are three possible outcomes of interest for proteins belonging to a given GO class according to these analyses. First, a given GO term can result statistically significant in the first test (disordered classes in *A. thaliana*) but not in the second one (comparison of with Human) if, for example, the disorder content of proteins is similar in both organisms. Second, a given class could be statistically significant in the second test but not in the first one. This would imply that while the disorder content of proteins annotated with that given functional class is not particularly high in *A. thaliana*, it is still significantly higher than the disorder content of the equivalent (i.e. annotated with same functional class) proteins in Human. Third, a given GO term could be present in both tests. This case corresponds to a class that is both significantly enriched in disorder in *A. thaliana* and has more disorder than its Human equivalent.

In both analyses, the set of significant GO terms reported by each test was further examined to reduce it to a smaller, more tractable set to be used for the biological interpretation of the statistical results. In order to accomplish this, we condensed our list of significant terms with the ReviGO tool[264]. This computational method "collapses" a set of GO terms based on several measures of semantic similarity by removing functional redundancies. The result is a smaller number of representative terms, which is easier to handle and interpret. These resulting terms correspond to the cluster representatives (which are graphically represented as a

single rectangle, see Figure 3-5 and Figure 3-4), and their choice is unaffected by whether the terms are more general or more specific. The size of each rectangle (cluster representative) represents the "uniqueness" of the term. This rectangle size assesses whether the term is an outlier when semantically compared to the whole list, in other words, it measures the frequency of the GO term in the underlying GO database[264]. The clusters' representatives are automatically joined into "superclusters" of loosely related terms identified with different colors. This representation allows a general visualization of the terms while discarding any overrepresentation of similar functional terms (Figure 3-5 and Figure 3-4).

3.4. Results

3.4.1. Overall disorder in *A. thaliana*

The analysis of overall disorder content of *A. thaliana* revealed that its proteome is, on average, less disordered than that of Human. Table 3.1 shows the different metrics of "disorder" in *A. thaliana* and *H. sapiens*. There were significantly more disordered proteins in human (defining "disordered protein" as one with ≥ 50% of disordered residues) with respect to *A. thaliana*: 35.9% vs. 29.5% (Figure 3-2; Chi-square test; p-value: <2.2E-16). The percentage of proteins with at least one "long disordered region" (LDR) was also higher in Human (68.5% vs. 57.2%, Chi-square; p-value: <2.2E-16;) and so was the average number of LDWs per protein (1.46 vs. 0.96; Wilcoxon Raked Sum test; p-value <2.2E-16). Furthermore, the average number of residues that fell into these LDWs was also higher in human

(27.0 vs. 19.7; Wilcoxon raked sum test; p-value <2.2E-16). Although these results are based on Disopred predictions, the tendency was also maintained for the other predictors and so was its statistical significance (see Appendix A, Table 3A).

Table 3-1. Summary of intrinsic disorder metrics for *A. thaliana and H. sapiens.* Results shown for Disopred (disorder prediction) and ANCHOR (Disorder binding regions, DBRs). For results obtained with other predictors see Appendix A, Table 3A.

Disorder metric	***A. thaliana***	***H. sapiens***
Mean content of disorder	29.5%	35.9%
Proteins with at least one LDWs	57.2%	68.5%
Mean number of LDWs	0.96	1.46
Mean number of residues belonging to LDW	19.67%	27.04%
Proteins with at least one DBR	50.7%	66.3%
Mean DBR per protein	2.34	5.11
Mean resides belonging to DBR	8.4%	13.8%

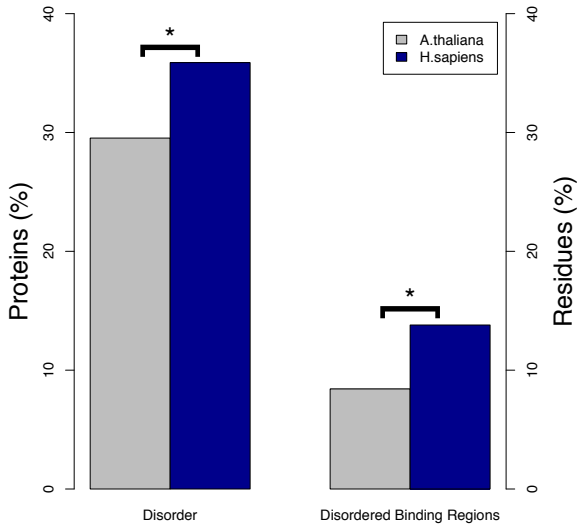

Figure 3-2. Overall predicted global disorder and disordered binding regions in *A. thaliana* and *H. sapiens* proteins. Left: percentages of disordered proteins (disordered proteins criterion: proteins containing at least 50% disordered residues based on Disopred predictions). Right: average percentages of disordered residues involved in binding (DBRs), as predicted by ANCHOR. The stars denote significant differences evaluated with the same Chi-square tests described in the Section 3.3.

In order to assess if the inter-species difference in disorder content is only observed for highly disordered proteins (≥ 50% of disordered residues) or if it is also observed in proteins with other ranges of sequence disorder, proteins were grouped according to the percentage of predicted disorder of their sequence (Figure 3-3A). The distribution was shifted to lower percentages of disorder (0-30%) in *A. thaliana*, while in Human it was shifted to higher disorder content (30-100%). These differences were statistically supported and predictor-independent (see

Appendix A, Figs. 1A, 2A, 3A), with the exception of the 30-50% bin for the VSL2 predictor, for which there was not statistical difference between both organisms.

The human proteome was also more enriched in predicted disordered regions potentially involved in protein-protein interactions (Disorder Binding Regions, DBRs). While 50.7% of *A. thaliana* proteins had at least one DBR, the proportion for Human was of 66.3% (Chi-square; p-value <2.2E-16). The average number of DBRs per protein was also higher in Human (5.11 vs 2.34, Wilcoxon Raked Sum test; p-value <2.2E-16). The average content of disordered-binding residues was higher in Human than in *A. thaliana*: 13.8% vs. 8.4% (Wilcoxon Rank Sum test; p-value <2.2E-16) (Figure 3-2 and Table 3-1). When proteins were grouped according to intervals of DBR residues content, there was always statistical difference between the number of DBR residues for both species, with more disordered binding residues in Human (Figure 3-3B).

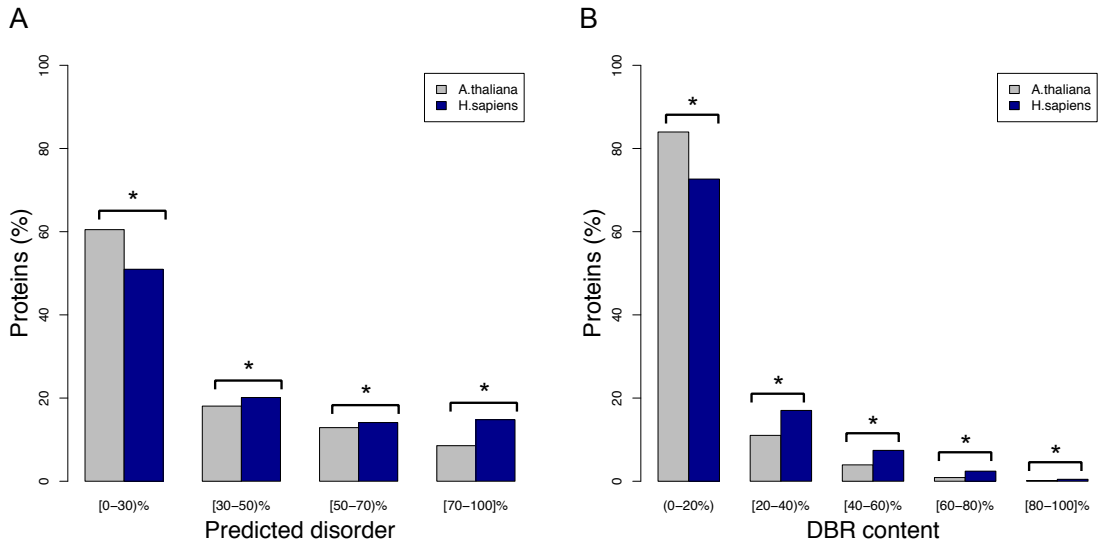

Figure 3-3. Fraction of proteins with different degrees of predicted disorder and disordered binding regions in *A. thaliana* and *H. sapiens* . A) Protein disorder (quantified as the percentage of disordered residues with respect to the sequence length) is binned into different ranges. The data reported is obtained using Disopred predictions. B) Percentage of disordered residues (calculated as reported in A) involved in binding predicted by ANCHOR. The stars denote significant differences evaluated with the same Chi-square tests described in the Section 3.3.

3.4.1.1. Protein disorder and functional categories

As described in the previous section, in the first part of this analysis we evaluated which functional categories were significantly enriched in disordered proteins in *A. thaliana*. In the second part, we performed a comparative analysis to detect functional classes that were distinctively associated to disorder in this organism with respect to Human. The complete set of GO terms resulting from

these two evaluations are shown in Appendix A Table 1A (*A. thaliana*), and Appendix A Table 2A (*A. thaliana* vs. Human).

3.4.1.2. Disordered proteins in *A. thaliana* are enriched in cell cycle, signaling and response to stimulus.

The list of significantly enriched GO terms from the disordered proteins of *A. thaliana* was analyzed using ReviGO in order to obtain a smaller set of representative terms that would facilitate its biological interpretation (for further details see 3.3.4). A schematic representation based on the ReviGO summarizing GO biological processes that were detected by DAVID as overrepresented (p-value ≤ 1E-5) in the set of disordered proteins of *A. thaliana* (those with at least one LDW according to Disopred predictions; Section3.3) is shown in Figure 3-4. The complete list of the GO terms is available in the Appendix A, Table 1.

Functional categories enriched in disordered proteins in *A. thaliana* (Figure 3-4) included "post-translational protein modification" (comprising nucleic acid metabolism, gene expression, protein synthesis and maturation) and a category labeled by ReviGO as "response to red or far red light". The latter, in addition to light signaling, included "response to endogenous and abiotic stimulus" and most of the hormonal signaling pathways. Therefore, this category could be better summarized as "response to stimulus". Other significantly enriched terms were "pattern specification", "transport"/"secretion", "cation homeostasis", "cellular compartment organization" (mostly referring to chromatin and nucleosome assembly), "cell

cycle", and "reproduction". Another interesting category that is enriched in disorder is "vesicle-mediated transport", which will be discussed in more details in the Chapter 4. Similar results were obtained with other disorder predictors and other disorder criteria (See Appendix A, Figures 4A-7A). In conclusion, these functional classes could be summarized as "signaling", "development", "cell cycle" and "response to stress" (light, abiotic, etc.), and they were represented mainly by proteins belonging to hormonal signaling pathways or transcription factors.

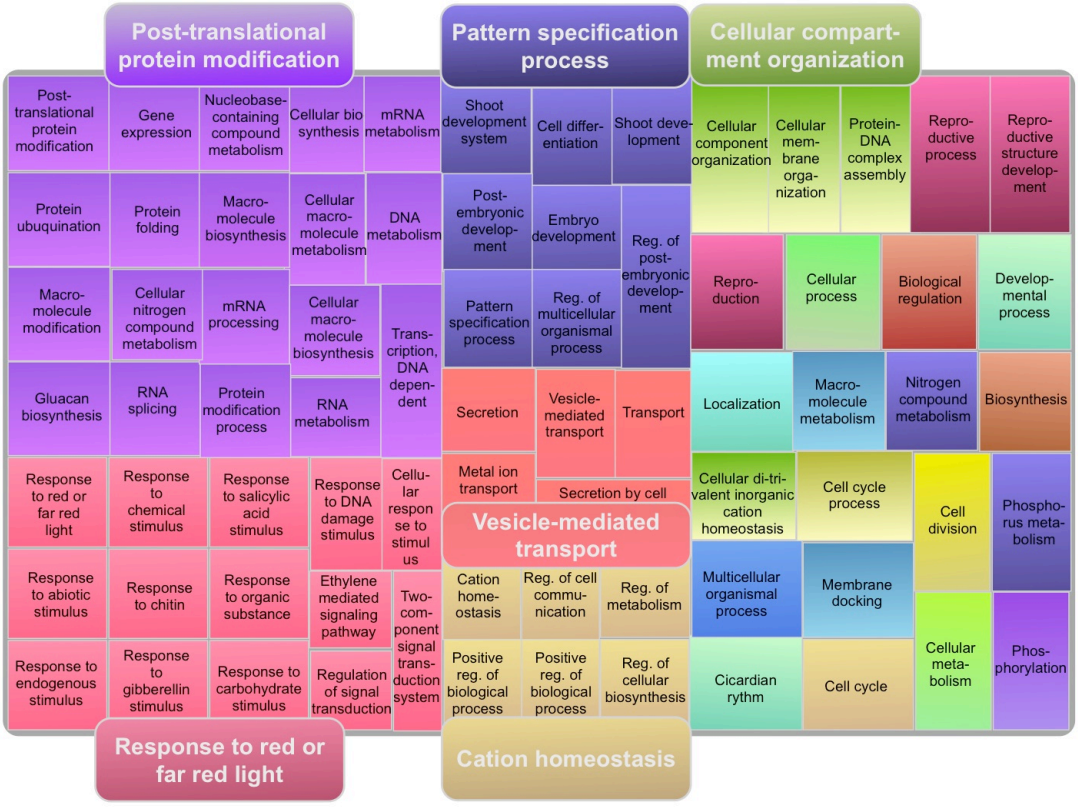

Figure 3-4. Representation of the main GO "Biological Processes" significantly enriched in disordered proteins in *A. thaliana*. Disordered proteins here correspond to those with one or more "long disordered regions" (LDR) based on Disopred predictions. This schematic representation was adapted from ReviGO, a method for summarizing and visualizing lists of GO terms. Each rectangle represents a cluster of related terms labeled according to a representative term. Rectangles are grouped in "superclusters" (identified with the same color) based on SimRel semantic similarity measure.

3.4.2. Disordered proteins in *A. thaliana* are more enriched in environmental detection and adaptation related functions than disordered proteins in *H. sapiens*.

A schematic representation based on ReviGO, summarizing the GO biological processes with a significantly higher proportion of disordered proteins in *A. thaliana* compared to Human (p-value ≤ 1E-5) is shown in Figure 3-5. As in the previous section, disordered proteins corresponded to those with at least one LDW according to Disopred predictions (Section 3.3). The complete list of terms is available in Appendix A, Table 2. While 146 GO terms were significantly enriched in disorder in *A. thaliana* (previous section), there were only 88 terms for which the disorder degree was significantly higher than in Human. Again, we found enrichment in categories associated to "detection and response to stimulus". In this case, however, most of such categories were related to external and xenobiotic stimulus (Figure 3-5).

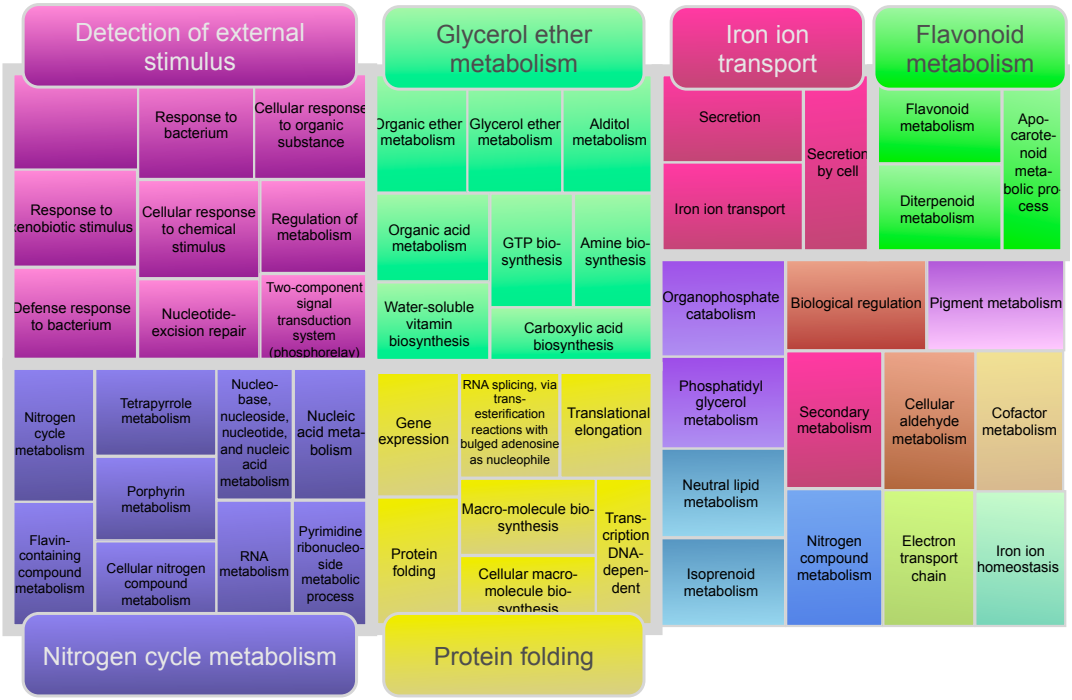

Figure 3-5. Representation of the main GO "Biological Processes" comparatively enriched in disordered proteins in *A. thaliana* with respect to *H. sapiens.* Disordered proteins correspond to those with 1 or more LDWs based on Disopred predictions. This schematic representation was adapted from ReviGO, a method for summarizing and visualizing lists of GO terms. Each rectangle represents a cluster of related terms labeled according to a representative term. Rectangles are grouped in "superclusters" (identified with the same color) based on SimRel semantic similarity measure.

A detailed view of the "response to stimulus" GO:BP subgraph, is shown in Figure 3-6, highlighting the terms which are enriched in disorder in *A. thaliana* (green nodes), as well as those more enriched in *A. thaliana* when compared to Human (blue nodes). It can be easily appreciated that the latter terms are more related to external stimulus.

The fact that the terms “response to endogenous stimulus”, “cell cycle”, etc. are no longer enriched indicates that proteins of these particular categories have similar disorder content in Human and *A. thaliana*. In contrast, “Protein folding” (including nucleic acid metabolism, gene expression, protein synthesis and maturation) was again present in the comparison between the two organisms, indicating that these processes are more disordered in *A. thaliana* than in Human. Other functional categories with significant disorder included those related to nitrogen metabolism and other molecules (flavonoids, glycerol, isoprenoids, cofactors, pigments). Other disorder predictors and disorder criteria provided similar results (See Appendix A Figures 4A-6A), especially for “detection of xenobiotic/external stimulus”, which repeatedly appeared as a functional category more disordered in *A. thaliana* than in Human, independently of the predictor and criteria used.

In conclusion, GO functional classes that are more disordered in *A. thaliana* compared to Human can be divided into two major related functions: “detection and signaling of external stimulus” (including chaperone activity induced by stress, related to “protein folding”) and “secondary metabolism”. In the case of plants, the latter is intrinsically related to the response to external stimulus, because plants have developed secondary metabolites as major tools to cope with environmental stress. Among proteins annotated under “detection of external stimulus” and “nucleotide-excision repair”, it was remarkable the high amount of those involved in

perception and signaling of light quality, which is the most influential external stimulus in plant development.

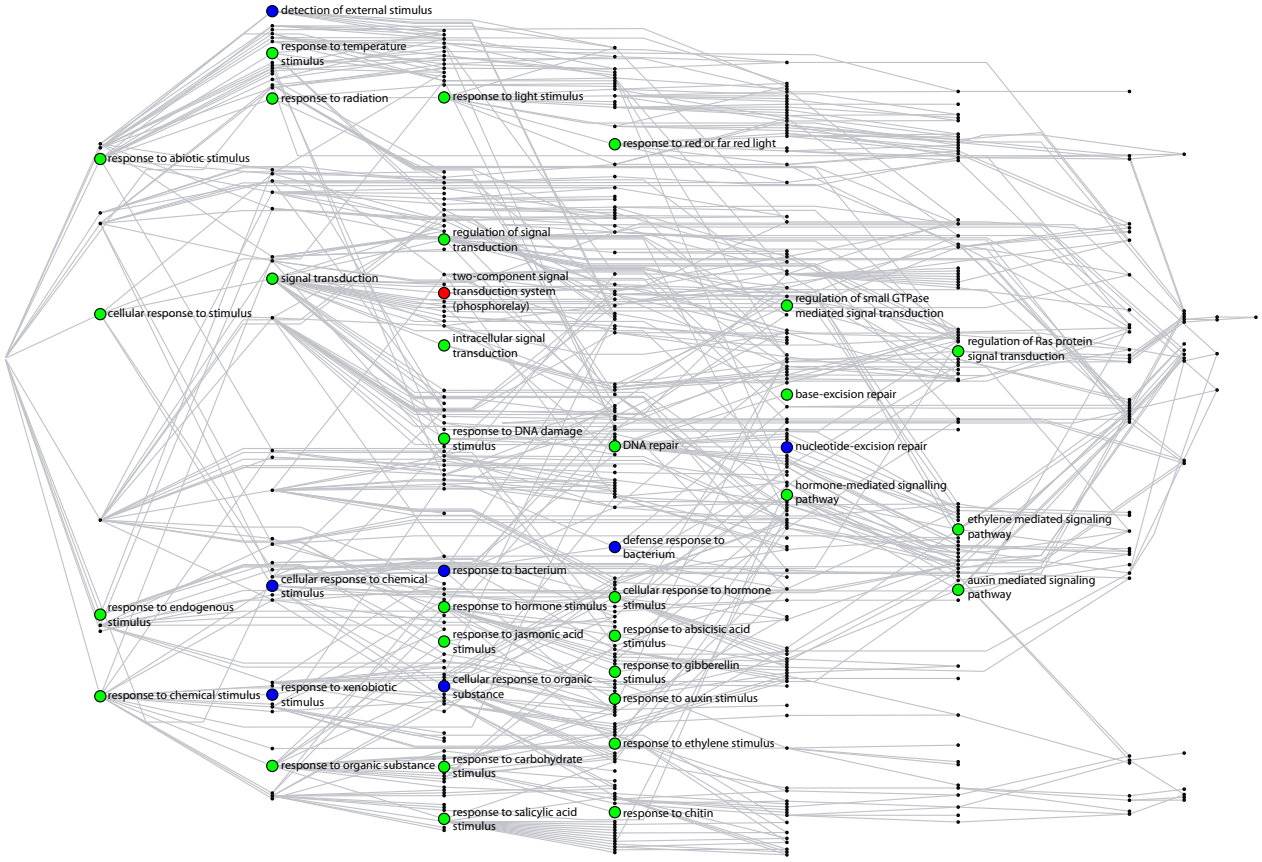

Figure 3-6. Subgraph of biological process "Response of stimulus" (GO:0050896). Green nodes correspond to those GO:BP terms significantly enriched in disorder in *A. thaliana*. Blue nodes correspond to those GO terms enriched in disorder in *A. thaliana* compared to Human. The red node represents the only common term between these two sets.

Finally, we conducted the same analysis to compare disordered binding regions (DBRs) in *A. thaliana* and human proteins. The functional categories we found are also related to "detection/response to xenobiotic/external stimulus", "defense against bacteria", "multi-cellular processes", and number of metabolic processes (See Appendix A, Figure 7A).

3.5. Discussion

The success of evolution in generating organism complexity has been paralleled by the increase in complexity in the underlying cellular and molecular processes. One can conceive two main ways for increasing the plasticity and complexity of a biological process supported by a network of protein-protein interactions: either increasing the number of proteins or increasing the number of interactions ("wiring"). Protein-protein interactions mediated by unstructured regions are recognized as a way of conferring plasticity to protein interaction networks[82,142,99]. Additionally, due to the physico-chemical characteristics of the interactions mediated by flexible and disordered regions, they are frequently involved in transient interactions with multiple partners[133]. Accordingly, increasing the disorder content of a particular subnetwork (biological process) of the interactome might be an efficient way to increase its "wiring", the possible connections between the proteins and, consequently, the plasticity of the system, without increasing the number of proteins involved.

Disorder content has been found to positively correlate with what is most commonly intuitively recognized as "organism complexity". The fact that the amount of IDPs and IDRs increases with organismal complexity is usually explained by the observation that more disorder is needed for signaling and coordination among the various organelles of eukaryotes compared to prokaryotes[11,69,70]. Although it is intuitively understood, a consensus on the formal definition of organismal complexity and its quantification is currently lacking. According to Shad and colleagues, for instance, complexity can be measured based on the number of different cell types in an organism ranging from 1 (in bacteria) to about 200 (in humans)[246]. A common approach is to apply the concept of complexity in a broader sense, where complexity is related to the amount of information an organism needs to function properly[124]. Most functions mediated by intrinsic disorder are in fact linked with complex responses to environmental stimuli and communication between cells, which has raised the question of whether intrinsic disorder can be linked to organismal complexity. Regardless of the strictness of its definition, organismal complexity is clearly a multiparametric trait that can be affected by several features, such as the organism's alternative splicing capacity, potential to interact with the environment, tissue-specificity and obviously, proteome size. Thus, it is crucial to correlate intrinsic disorder also with the complexity of the system being examined rather than just correlate it to the complexity of the organism itself.

Our results show that, as expected, the human proteome is globally more disordered than *A. thaliana* proteome. This trend is observed regardless of the predictor and criterion used for defining disorder. At the same time, both Human and *A. thaliana*, as complex eukaryotes, are also much more disordered than bacteria[171,246,68].

In agreement with previous observations[142,16,75,10], we found that disorder in *A. thaliana* s is involved in biological processes rich in transient interactions with multiple partners (e.g. cell cycle, signaling, and DNA and RNA metabolism, including splicing). These processes, which are generally more complex in eukaryotes than in prokaryotes or that may even represent new acquisitions of evolution (e.g. splicing), are the prototypical processes that have been previously related to disorder in higher organisms.

It is striking, however, that despite the fact that all intrinsic disorder criteria evaluated in this study point to higher disorder levels for Human – in accordance to its higher complexity – we find some functional classes for which disorder is significantly higher in *A. thaliana.* More interestingly, these GO classes are related to processes such as environmental perception and response – for which plants have developed more complexity – and are fundamental for their adaptation. Our results support previous evidence of the relationship between intrinsic disorder and processes related to the response to environmental stimuli[246, 124,125, 126].

The ability to accommodate its phenotype to changing environmental conditions, or *phenotypic plasticity*, is very important for adaptation and survival of

any organism. Plasticity is particularly relevant in plants since these sessile organisms cannot escape from environmental challenges as animals can do[20]. Specifically, plant plasticity depends on the capacity to identify the challenge, integrate the external information through signaling pathways, and finally change the basal developmental programs to stress programs (which include the production of secondary metabolites) to adapt and survive to those threats.

There are processes for which plants have developed particularly complex mechanisms. Light, for instance, is probably the most important external clue for plant development, and plants have developed complex perception (photoreceptors) and signal transduction mechanisms to finely tune their growth and development according to light quality and intensity[265,266,267,268]. A recent study[249] (published after this analysis was concluded) highlights the importance of intrinsic disorder in plant chloroplasts. Remarkably, "response to light stimulus" appears as one of the most disordered GO terms in *A. thaliana* and also in the comparison of *A. thaliana* to Human. Other two GO terms enriched in disordered proteins in *A. thaliana* as compared to Human are "detection of external stimulus" and "nucleotide-excision repair". The first one annotates many proteins involved in perception and response to dark and Red/FarRed light (COPs, SPAs, PIFs, etc.). The second GO term includes proteins involved in UV light perception and response. Thus, processes for which plants have developed mechanisms more complex than humans appear more disordered, further supporting the correlation between complexity – in this case of a given set of functions – and intrinsic disorder.

Another example of complexity in plant development and function is their ability to adapt to abiotic stress, such as drought, salinity or the cell desiccation that occurs during seed development. Consistent with this complexity, several GO terms enriched in disordered proteins (compared to Human) are related to protein folding or abiotic-stress related signaling. Moreover, among the few plant proteins for which disorder has been previously described, ERD10 and ERD14 are examples of chaperones whose structural disorder provides the flexibility to interact with many different partners and prevent their denaturation and aggregation[253].

An additional set of GO terms significantly disordered in *A. thaliana* as compared to Human is related to secondary metabolism ("flavonoid", "isoprenoid", "pigment", "nitrogen", "vitamin", "cofactor", etc.), which use, in many cases, represents evolutionary acquisitions of plants to cope with environmental stress and adaptation. Some flavonoids and anthocyanins, for instance, are produced by plants to protect from UV radiation (another GO term more disordered in *A. thaliana*, as reported above), whereas other secondary metabolites are involved in attracting pollinators or defending from predators[269,270,271]. In the case of nitrogen, it is often a limiting factor for plant growth. Multiple nitrogenous compounds are involved in different functions in plants, including storage of nitrogen, but they are also related to defense and signaling[271].

It was proposed that "increasingly integrating protein disorder into the toolbox of a living cell was a crucial step in the evolution from simple bacteria to complex eukaryotes"[124]. Our results support the correlation between organism

complexity and protein disorder, and suggest that plants have used disorder as an evolutionary tool to increase complexity in their biological/protein networks. This increased complexity is particularly evident in those networks underlying phenotypic plasticity and adaptation to environmental stress.

In conclusion, the genome-wide analysis of intrinsic disorder in *A. thaliana* reported in this study enabled the identification of functional classes that are enriched in disordered proteins. In addition, the functional classes identified contained proteins involved in processes related to this *A. thaliana* adaptation and response to the environment. This phenomenon perfectly fits the notion that newly introduced disordered proteins and protein segments mainly serve as carries for new binding regions in eukaryotic organisms[85], and thus add complexity to the system. Thus, our results provide compelling evidence to demonstrate that intrinsic disorder provides a mechanism to increase *Arabidopsis thaliana*'s ability to adapt to the environment.

Chapter 4

Role of intrinsic disorder in cellular functions: Analysis of intrinsic disorder in proteins involved in the human and yeast vesicular trafficking machineries

4.1. Introduction

Vesicle trafficking systems provide a mechanism for communication between different intracellular compartments and between the cell and the extracellular space. Cargo molecules, including proteins, lipids and signaling molecules, are transported via one of the three major vesicle trafficking routes to their final destinations. Trafficking in each route is mediated by vesicles containing specific coat protein complexes. Clathrin-coated vesicles mediate endocytosis and the late secretory route, while the coat protein complex II (COPII) and coat protein complex I (COPI) vesicle trafficking routes are responsible for the bidirectional traffic

between the ER and the Golgi apparatus. Clathrin, COPI and COPII mediated routes – ubiquitous in eukaryote cells – not only deliver cargo molecules to the plasma membrane and to specific organelles, but are also responsible for maintaining the physiologic protein and lipid composition in intracellular organelles.

Despite the similar fundamental organization in regulatory mechanisms and structural features of these three systems, the molecular machinery, functions and evolutionary characteristics differ significantly. Some of these functional and evolutionary differences have been previously studied. However, the structural features that mediate these differences remain uncharacterized. Evidence showed that long regions of certain clathrin-associated adaptor proteins are disordered, unstructured or unfolded[272,273,274,275]. Furthermore, we demonstrated that a number of proteins related to vesicle-mediated transport in *A. thaliana* are highly disordered (Chapter 3). We hypothesize that disordered proteins play a key role in vesicle trafficking, which, in turn, could explain some of the functional and evolutionary differences among transport routes. In this study, we aimed to investigate the location of disordered regions within proteins involved in vesicular traffic and their function.

Although the protein machinery of the clathrin, COPI and COPII mediated routes differ almost completely, these routes share several main structural and mechanistic characteristics. In all three routes, *i) a specific multisubunit protein coat complex on the outer surface of the vesicles self-assembles as a lattice* collecting and concentrating the appropriate adaptor-cargo complexes into membrane patches.

This assembly process is responsible for cargo selection and for enhancing the budding and fission of the vesicles (by assisting membrane curvature generation[276]). Proteins involved in the assembly of the cage components have a common underlying structural design. They all have N-terminal β-propeller containing WD40 structural motifs and several α-solenoid motifs towards the C terminus[276]. The three routes also *ii) depend on different small GTPases for coat assembly*, as well as on the corresponding activating or nucleotide exchanging factors[277]. In all three routes, *iii) vesicles are uncoated after formation* (they are stripped of both the cage-forming scaffold proteins and the adaptor proteins). Additionally, *iv) the process of cargo handling* is also shared among the three routes. Adaptor proteins link the scaffold to the cargo and to the membrane and communicate with proteins involved in the formation and fission of the transport vesicle[278,279]. Cargo-specific receptor proteins are also present in all three systems. Other similarities between these routes include *v) the basic mechanisms of vesicle transport* (driven by motor proteins along the actin cytoskeleton elements) and *vi) the mechanism of vesicle fusion into the target membrane*. The key players of vesicle fusion are the SNARE (Soluble N-ethylmaleimide-sensitive factor attachment protein receptor/SNAP receptor) proteins. SNAREs generate the pulling force required for placing the two membranes in proximity for fusion[280,281]. Finally, *vii) membrane fusion regulation* is also common between the different systems. For example, multisubunit tethering complexes or coiled-coil tethers[282] help COPI and

COPII vesicles getting close to the target membrane so that SNARE proteins can interact and arrange fusion[283,284].

Despite this array of mechanistic, structural, and regulatory similarities, there are fundamental functional and evolutionary differences between the clathrin-mediated system and the other two systems. While COPI and COPII vesicles mediate ER-Golgi-ER trafficking and are essential for cell viability[285,286], the clathrin-vesicles seem less indispensable. Knock-outs of clathrin components, such as AP-2 (Adapter-related protein complex 2) are lethal in multicellular organisms[287]. However, yeast cells with chromosomal deletions in gene encoding AP-2[288] or even clathrin[289], but not of other clathrin pathway associated adaptors (e.g. epsin-homologs[290], HIP1; Huntingtin-interacting protein 1, and Hip1R; Huntingtin-interacting protein 1-related protein[291]) are viable. These observations suggest that clathrin-mediated routes present higher structural and functional plasticity and robustness[288,292] than COPI and COPII mediated systems, which from an evolutionary point could translate into more adaptability.

These three trafficking pathways also differ in their complexity. In the COPI and COPII systems, the adaptor subunits are part of the multisubunit coat complex and there is only one set of subunits for each pathway[293]. However, some of the subunits of both COPI and COPII systems have different isoforms with different cargo specificity or differential localization, suggesting distinct functional roles for the coat complexes[294,295]. Adaptor proteins from clathrin-mediated system, on the other hand, comprise a highly diverse and dynamic set of proteins. These proteins

may share similar functions (e.g. binding the clathrin coat and the cargo at the same time), or play individual roles in the assembly and transport of vesicles[279,296]. Adaptor proteins can participate in different routes, showing preference for different sorting signals and organelle membranes. In addition, these proteins may sort different cargo types into the same population of vesicles cooperatively, or recruit cargo into different populations of vesicles on the same membrane in a mutually exclusive manner. Clathrin adaptor proteins, are key players in the assembly of the large, highly complicated macromolecular complexes and usually interact with many interaction partners[296,297].

We hypothesized that intrinsic disorder provides functional and structural advantages for vesicle trafficking proteins. Disordered regions, for example, could be mediators of the exceptional diversity, plasticity or adaptability of clathrin pathway adaptors. Given their enlarged capture radius, disordered regions could offer the ability to bridge large distances via the "fly-casting mechanism" of protein binding[96], thereby promoting effective assembly of the vesicle coat, as described in Section 1.2.1. Since short linear protein interaction motifs[298,299], posttranslational modification sites[300,90], and tissue-specific disordered binding regions of splice variants[184,301] usually reside in disordered protein segments, these regions could be especially important in mediating specific binding to partner proteins and in displaying important regulatory roles[279,274]. All the above mentioned characteristics of disordered regions, along with other advantages they provide – such as their conformational freedom and their ability to bind many interactions partners (i. e.

moonlighting[74]) – make disordered regions excellent candidates for the binding involved in the assembly and transport of macroscopic organelles[302].

To the best of our knowledge, the abundance of disordered regions was only assessed for proteins in the clathrin pathway by secondary structure prediction methods[274], which do not allow proper identification of structurally disordered protein segments. Furthermore, the presence of structural disorder was not addressed in either COPI nor COPII mediated systems. Thus, a quantitative assessment of protein disorder content in these systems using adequate methods is still lacking. We present a systematic comparative study of protein disorder in all three main vesicle trafficking systems using protein disorder prediction methods. The quantification of intrinsic disorder in proteins involved in the different vesicle trafficking pathways, together with a systematic comparison of their disorder content aided in understanding how the structural properties of these proteins affect their functional and evolutionary features.

4.2. Hypothesis

Hypothesis: *Intrinsic disorder may be responsible for some of the functional and evolutionary differences present in the main vesicle trafficking routes*.

4.3. Methods

To assess the intrinsic disorder content of proteins involved in the main vesicle trafficking routes, we compiled datasets of human and yeast proteins

involved in the clathrin, COPI and COPI mediated routes and classified them according to their functional roles. We used the human and yeast proteomes as background datasets. We assessed the intrinsic disorder content for each protein sequence in each dataset using different disorder metrics. Then, for proteins belonging to the vesicle trafficking dataset, we compared the intrinsic disorder content in the different functional groups and in the three vesicle trafficking routes. In addition, we compared disorder content across the two organisms and against the background datasets. Finally, we investigated single cases in which disordered regions of human proteins involved in vesicle trafficking seemed crucial for the protein's function and inspected their orthologous proteins in yeast.

4.3.1. Datasets of human and yeast proteins involved in vesicle trafficking systems

An extensive literature search was conducted to collect coat proteins, adaptors and the most important enzymes of the clathrin-mediated route[278,296] and the main components of the COPI and COPII vesicle coats[285,286,303]. Members of the different clathrin-independent endocytosis routes and proteins involved in the vacuolar traffic are poorly characterized to date[304] and were not included in this study. Proteins involved in vesicle fusion regulation (multisubunit tethering complexes, coiled coil tethering proteins, SM (Sec1/Munc18-like) proteins, and other regulatory proteins) were also compiled[282,284,305]. The main mechanistic promoters of vesicle fusion – the SNARE proteins – were similarly collected[280]. All proteins involved in regulation of vesicle fusion and all SNARE proteins were

collected along with their corresponding functions, as they present similar functions in the three trafficking routes[284].

The datasets (for both yeast and human proteins) were extended by adding interaction partner proteins reported to belong to any of the three main systems in Universal Protein Knowledgebase database[306] (UniProtKB release 2012_09). Additionally, we inspected proteins annotated with vesicle trafficking-related terms of the Gene Ontology (GO) Database[231] (namely GO:0048208, GO:0012507, GO:0006892, GO:0030126, GO:0030130, GO:0030132, GO:0030136 GO:0048205, GO:0006890). Only proteins taking part in one of the main vesicle trafficking routes according to their UniProt annotation or the literature were included. The resulting datasets (244 human proteins; 162 yeast proteins) contain only manually curated entries (see Appendix B, Table 1B for human proteins and Table 2B for yeast proteins).

4.3.2. Human and yeast protein background datasets

The protein sequences from the complete human and yeast proteomes were extracted from the UniProtKB database (release 2012_09). The queries specified organism (yeast or Human) and included the complete reviewed proteome set (keyword 181). Protein sequences were additionally filtered for fragmented proteins, and the resulting datasets were used as the background datasets.

4.3.3. Functional classification

Proteins were classified according to their functional roles as reported in published literature. The protein dataset was divided into seven functional groups. Four groups include budding and fission-associated proteins: i) coat proteins, ii) adaptors and sorting proteins, iii) enzymes and enzymatic activity related proteins, and iv) unclassified proteins, which includes all the proteins that could not be classified into the first three budding and fission-associated groups, many of them being transmembrane cargo-specific receptors. These four functional groups of proteins were subclassified according to the three main vesicle trafficking systems (Clathrin, COPI and COPII mediated). The other three functional groups include fusion-associated proteins: v) SNAREs; vi) multisubunit tethering complexes; and vii) other fusion regulators. In the human proteins, a functional group of fusion regulator proteins that play a specific role in neurotransmitter transport was also added.

4.3.4. Identification of transmembrane segments and Pfam protein domains

The location of transmembrane regions in the protein sequences of vesicle trafficking protein datasets was assigned according to the annotation in the UniProtKB. Additionally, protein domains and their corresponding locations were assigned using the PfamScan method to scan the Pfam protein families database[307]. From this database, only the Pfam-A (v26.0) entries were used, corresponding to

manually curated protein families. Default domain coordinates were assigned using alignment coordinates provided by the HMMER3 tool based on Pfam-A HMM profiles for the search[307].

4.3.5. Prediction of protein disorder

The prediction of intrinsic protein disorder was carried out using ad-hoc scripts based on the prediction methods IuPred and ANCHOR as described in Section 3.3.3. For each protein in the two datasets (Human and yeast), the previously defined standard measures used to describe the disorder content of proteins were calculated: i) relative disorder content, ii) number of long disordered regions (LDR) for different k lengths ($k \geq 30$, 50 and 100 residues) and ratio of residues in LDR, and iii) disordered binding regions, DBRs and ratio of residues in DBRs. For proteins in the vesicle trafficking datasets with transmembrane regions, residues belonging to transmembrane segments were not taken into account for the calculation of any disorder metric, since disorder predictors might identify these sequences as disordered regions.

4.3.6. Identification of orthologous proteins

Orthology identification between human and yeast proteins in the vesicle trafficking dataset was performed using the InParanoid[308] tool (v7). This tool uses pairwise similarity scores (calculated using NCBI BLAST) between two complete proteomes for constructing orthology groups. Each orthology group is composed of two seed orthologous proteins (one from each proteome) and of any sequence in

either proteome that is closer to the corresponding seed ortholog than to any other sequence in the other proteome.

4.3.7. Identification of protein complexes involving disordered protein segments

A comprehensive search in the Protein Data Bank[309] (PDB) was performed to identify complexes of distinct pairs of vesicle trafficking-related proteins in which the binding region of at least one of the partners is predicted to be intrinsically disordered by IuPred.

All data processing in this study was performed using ad-hoc scripts written in Perl programming language. All analyses were implemented in the statistical analysis programming language R[263].

4.4. Results

4.4.1. Classification of proteins involved in vesicle trafficking pathways

We assembled a comprehensive dataset of proteins involved in the three main vesicle trafficking systems in Human and yeast. To the best of our knowledge, this dataset, containing 244 human and 162 yeast proteins, is the largest and most complete collection of human and yeast proteins involved in vesicle trafficking. Each protein was identified by name and UniProt accession number, and classified according to the functional classification scheme (Section 4.3.2.) (for the Human and

yeast protein sets, see Appendix B, Table 1B and Table 2B). Disorder measurements and the location of transmembrane segments and Pfam domains are reported for each protein.

4.4.2. Human and yeast proteome sequences

The background datasets containing to the complete human and yeast proteomes comprise in 20,213 and 6,621 proteins.

4.4.3. Intrinsic disorder in human and yeast proteins

The number of proteins and the means and medians of all the disorder metrics calculated for each functional class and vesicle trafficking pathway in Human and yeast are summarized in Table 4-1. The mean and median ratio of residues in transmembrane segments and Pfam domains for the same groups are reported in Table 4-2. The number of proteins having long disordered regions of various lengths ($k \geq 30,50,100$) for the complete human and yeast proteomes, as well as for proteins in the different functional classes and vesicle trafficking pathways are reported in Table 4-3.

Overall, proteins involved in the three major vesicle trafficking pathways tend to be more disordered in Human than in yeast. In fact, 27% of human proteins and only 22% of yeast proteins present disorder content of at least 30%, which is considerably higher than mean ratio of disordered residues for proteins the background sets of either species (see below). The mean ratio of disordered

residues in human proteins (20.85%) significantly differs (Wilcoxon Rank Sum Test, p-value = 2.67E-02) from that of yeast proteins (17.77%). The difference in the mean number of DBR residues (Wilcoxon Rank Sum Test, p-value = 1.80E-02) is also significant. In Human, 44% of the proteins have at least one LDR (of length at least 30 residues), as opposed to 37% in yeast proteins having at least one LDR.

Table 4-1. Disorder content of proteins in the different functional groups of the three membrane trafficking pathways for Human (H) and yeast (Y). Proteins were classified in trafficking pathways are: Clathrin coat , COPI (coat protein complex I) and COPII (coat protein complex II) mediated pathways. Functional groups: COAT (coat associated proteins), ASP (adaptors and sorting proteins), EARP (enzymatic activity related proteins), UCP (unclassified proteins), MSTC (multisubunit tethering complexes), OFRP (other fusion regulatory proteins), SNARE (SNARE proteins) and NTSR (neurotransmitter transport specific regulators). For whole list of proteins, see Appendix B Table 1B (Human) and Table 2B (yeast).

	Number of proteins		Disordered residues (%) mean / median		Residues in Disordered Binding Regions (%) mean / median		Residues in Long Disordered Regions (%) (k ≥30 residues) mean / median	
Functional groups	H	Y	H	Y	H	Y	H	Y
COAT	10	7	22.76 / 9.20	19.50 / 7.58	15.35 / 6.06	11.01 / 4.00	18.06 / 4.67	14.51 / 5.08
ASP	64	38	25.17 / 21.49	25.20 / 15.80	14.80 / 8.93	13.85 / 8.22	17.28 / 8.13	18.16 / 8.54
EARP	18	9	24.88 / 22.84	20.59 / 10.68	15.49 / 14.95	10.77 / 4.02	16.64 / 14.37	14.30 / 0
UCP	32	32	16.77 / 6.95	14.11 / 5.53	9.35 / 0.96	5.83 / 0	7.29 / 0	7.86 / 0
MSTC	44	42	6.01 / 4.75	8.96 / 5.08	2.39 / 0.74	3.58 / 0.74	1.94 / 0	4.12 / 0
OFRP	19	16	17.82 / 12.86	10.74 / 7.88	7.89 / 5.93	3.67 / 1.65	8.12 / 0	4.44 / 0
SNARE	37	24	23.74 / 18.26	26.97 / 24.63	9.31 / 6.91	13.58 / 11.51	5.72 / 0	8.93 / 0
NTSR	25	-	31.74 / 19.43	-	14.79 / 8.44	-	20.75 / 8.65	-
Pathways								
Clatrhin	71	38	27.98 / 23.33	27.84 / 22.58	17.19 / 11.58	15.58 / 11.50	19.40 / 13.35	20.30 / 10.41
COPI	22	16	13.33 / 6.90	13.92 / 8.22	6.11 / 0.83	6.28 / 0.95	6.80 / 0	8.13 / 0
COPII	31	32	17.52 / 6.78	14.07 / 6.25	10.45 / 2.88	2.88 / 6.08	9.42 / 0	8.45 / 0

Table 4-2: Ratio of residues in transmembrane segments and Pfam entities (domains, families, repeats and motifs) for the different functional groups of the three membrane trafficking pathways for Human (H) and yeast (Y).

	Residues in transmembrane segments (%) mean / median		**Residues in PFAM entities (%) mean / median**	**Residues in PFAM entities (%) mean / median**
Functional groups	H	Y	H	Y
COAT	0 / 0	0 / 0	68.24 / 73.72	68.45 / 76.77
ASP	0 / 0	4.23 / 0	61.87 / 65.71	53.19 / 58.99
EARP	0 / 0	0 / 0	58.94 / 62.96	63.33 / 63.43
UCP	10.09 / 9.22	11.33 / 9.66	77.19 / 86.70	39.39 / 45.30

The average ratio of disorder content of human proteins involved in trafficking pathways (20.85%) is not significantly different from the average ratio disorder content of the complete human proteome (22.81%) as assessed by the Wilcoxon Rank Sum Test. Similar results were observed for yeast proteins, where the average ratio of disorder content of proteins in the trafficking pathways (17.77%) is not significantly higher than the average ratio of the whole proteome (16.96%).

Table 4-3: Number of proteins with disordered regions of various lengths for the different functional groups of the three membrane trafficking pathways for Human (H) and yeast (Y). Functional groups and pathways are defined as in Table 4-1. The last row refers to the whole proteome. Number of proteins with Long Disordered Regions of at least *k*=30,50 and 100 consecutive residues.

	Number of proteins		Proteins with LDR (≥30 residues) (%)		Proteins with LDR (≥50 residues) (%)		Proteins with LDR (≥100 residues) (%)	
Functional groups	**Human**	**Yeast**	**Human**	**Yeast**	**Human**	**Yeast**	**Human**	**Yeast**
COAT	10	7	60.00	57.14	50.00	57.14	30.00	28.57
ASP	64	38	60.94	57.89	45.31	44.74	37.50	28.95
EARP	17	8	70.59	50.00	58.82	50.00	52.94	50.00
UCP	28	27	28.57	22.22	21.43	22.22	14.29	11.11
MSTC	44	42	27.27	23.81	2.27	16.67	2.27	11.90
OFRP	19	16	36.84	37.50	26.32	18.75	26.32	0.00
SNARE	37	24	24.32	33.33	5.41	12.50	0.00	4.17
NTSR	25	-	60.00	-	40.00	-	12.00	-
Pathways								
Clatrhin	71	38	66.20	60.53	50.70	50.00	42.25	34.21
COPI	22	16	36.36	31.25	18.18	31.25	9.09	12.50
COPII	31	32	25.81	15.63	12.90	15.63	6.45	6.25
Proteome								
	20213	6621	45.60	35.45	33.11	24.51	18.18	12.49

4.4.4. Intrinsic disorder in protein functional groups

The different functional groups (the functional classification is described in Section 4.3.2) of both yeast and Human showed similarities in their disorder content (Table 4-1). A standard boxplot representation of the disorder content for the different functional groups is shown in Figure 4-1.

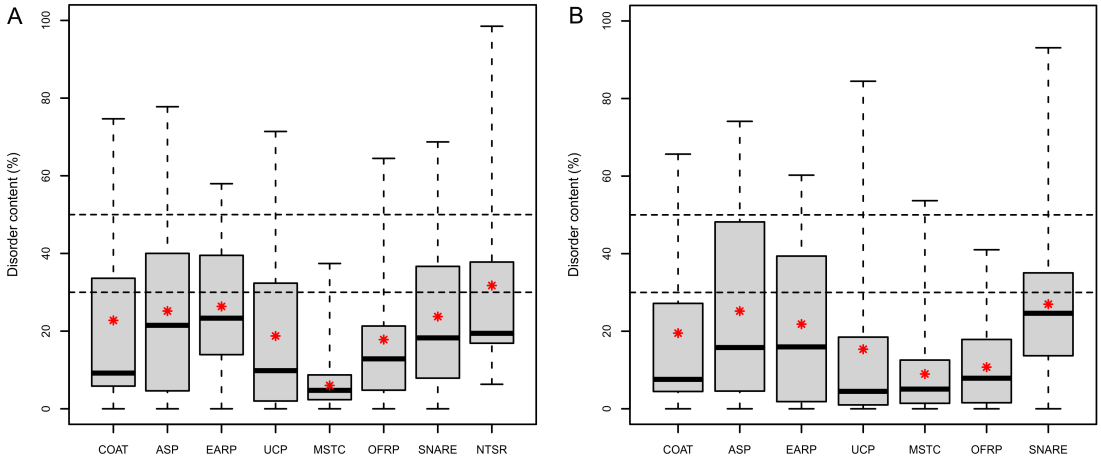

Figure 4-1 Disorder content for functional groups of proteins involved in vesicle trafficking. Fraction (%) of predicted disordered residues (disorder content) calculated using IuPred is presented for Human (A) and yeast (B) for data reported in Table 4-1. Functional groups are defined as in Table 4-1. The mean is depicted by a star. Proteins with disorder content (dc) (30%≤dc<50%) are considered fairly disordered; proteins with (dc≥50%) are considered highly disordered. The bottom and top of the boxes represent 25% and 75% of the data respectively, while the bold line in the middle of each box corresponds to the median (50%). The whiskers of the boxplots correspond the minimum and maximum values in the data, while the mean is depicted by a star.

Proteins involved in fusion regulation were among the least disordered, with a median disorder content of 4.78% (Human) and 5.08% (yeast) for the "multisubunit tethering complexes" (MSTCs) group, and 12.86% (Human) and 7.88% (yeast) for the "other fusion regulatory proteins" (OFRPs) group. Neither the MSTCs group (in Human) nor the OFRPs (in yeast) contained highly disordered (disorder content, d.c. > 50%) proteins, and their protein members are also largely depleted in LDRs. Similar results were obtained for the DBR residues of these

groups, where proteins in the MSTCs group have the lowest values, followed by protein in the OFRPs group.

Proteins in the “coat” and in the “unclassified” groups (the latter containing many transmembrane cargo-specific adaptors) are also rather structured in both species. However these two groups showed larger deviations in disorder content than the MSTC and the OFRP groups. Even in coat proteins, which form a completely folded, rigid cage-like structure on the surface of the vesicles, some of the subunits were predicted to be largely disordered. For example, disorder content of the clathrin light chains is 65.7% in yeast and 60.08% and 74.67% for human clathrin light chains A and B, respectively. Although to a lesser degree, the Sec31 (Protein transport protein Sec31) subunit of the COPII type coat is also considerably disordered (44.23% in yeast; 33.61% and 27.40% for human A and B paralogs, respectively).

The “SNARE” group is composed of proteins that belong to the same large protein class, all of them containing at least one v- or t-SNARE coiled coil homology domain and various types of family-specific domains. The proteins in the SNARE group show a surprisingly high variability in disorder content with a few of the proteins being mostly disordered and others being well structured. The median disorder content in human proteins is 18.26%, while in yeast dataset the group of SNAREs shows the highest median disorder content (24.63%) among all the functional groups common to both species. The member of the SNARE group with highest disorder content is the transport protein Sec9, which is almost entirely

disordered (93%). In addition, members of the syntaxin family of SNAREs have disordered N-terminal regulatory regions, which are involved in the interaction with SM proteins. Some of the interactions between syntaxin proteins and SM proteins have been extensively studied[310,305,284]. The disordered syntaxin N-tail folds into an ordered structure upon binding to it globular SM partner[310]. Different complexes between SNARE proteins of the sintaxin-family and SM-proteins that regulate the SNARE complex assembly were reported (Figure 4-2)[311,312,313]. The N-terminal regions of these SNAREs are predicted to be very disordered (at least 50% disorder content as predicted with IuPred) in their unbound form. In Figure 4-2, each interaction pair is represented by a PDB structure (left) and a domain map (as predicted by PfamScan) of the entire protein chain for each interaction partner (right). The disorder predictions obtained from this study are in agreement with previously reported evidence demonstrating that monomeric SNARE motifs are unstructured and form four-helix bundles only upon vesicle fusion[280,281].

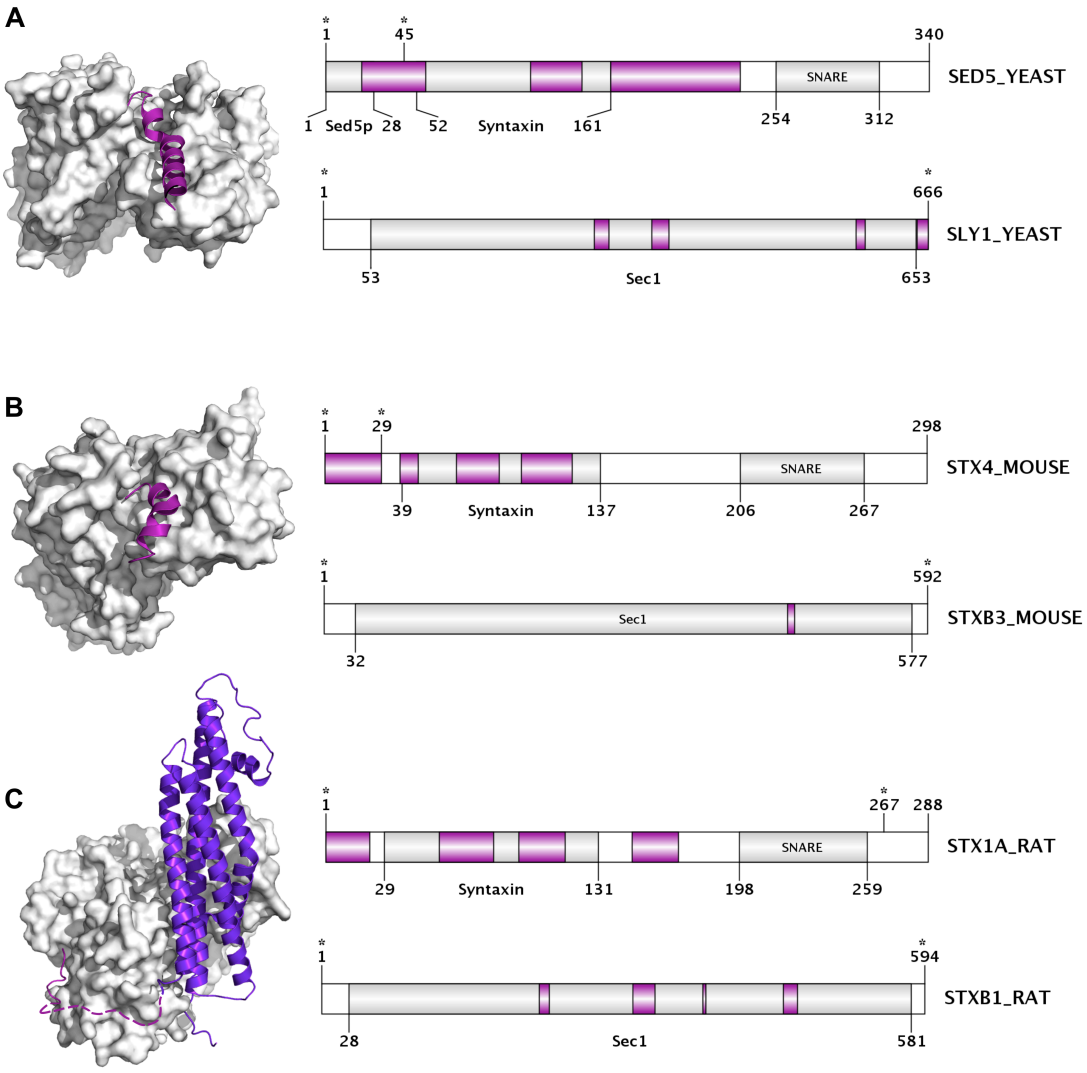
A
1
45
340
SNARE
SED5_YEAST
1
Sed5p
28
52
Syntaxin
161
254
312
1
666
SLY1_YEAST
53
Sec1
653
B
1
29
298
SNARE
STX4_MOUSE
39
Syntaxin
137
206
267
1
592
Sec1
STXB3_MOUSE
32
577
C
1
267
288
SNARE
STX1A_RAT
29
Syntaxin
131
198
259
1
594
STXB1_RAT
28
Sec1
581

Figure 4-2. Interactions between disordered N-terminal segments of SNARE proteins and folded SM protein partners. The N-terminal of the SNARE partner is predicted to be mostly disordered (disorder content ≥50%) in the unbound form. (A) Interaction of yeast syntaxin-family SNARE Sed5 and SM protein Sly1 (PDB: 1MQS). (B) Interaction of syntaxin-4 and syntaxin-binding protein 3 from mouse (PDB: 2PJX). (C) Interaction of syntaxin-1A (structure lacking the C-terminal transmembrane region) and syntaxin-binding protein 1 from rat (PDB: 3C98). The disordered SNARE N-terminal tails are represented with cartoon style (magenta) while the partner molecule is in surface representation (white) in the structures. In panel C, the remaining segment of syntaxin-1A (not part of the disordered N-terminal tail) is colored purple-blue, and the disordered residues of the N-terminal that are not included in the X-ray structure (10-26) are represented by a dashed-line. The domain map of the SNARE protein (top) and of the SM partner (bottom) are reported for each complex. Domain maps for each protein show names and locations of their reported Pfam domains. Disordered regions (length ≥ 3 residues) as predicted by IUPred are colored in magenta, while the structured segments are light-gray (predicted to be part of a Pfam domain) or white (if not predicted to be part of a Pfam domain). Regions present in the PDB structures are marked by stars.

The group of "adaptor and sorting proteins" (ASP) contained the highest number of extremely disordered (>50%) members in both the human and yeast datasets. Their median disorder content (21.49% Human, 15.80% yeast) however, is only the second largest in both species; proteins in the ASP group are very diverse in terms of intrinsic disorder content. It contains completely structured subunits of larger adaptor complexes (such as the sigma and mu subunits of the AP complexes, the zeta subunit of the COPI coatomer complex, and the Sec23 subunit of the COPII coat adaptor) and also mostly disordered (d.c. > 52%) adaptor proteins such as the epsins, DAB1 and DAB2 (Disabled homolog 1 and 2), HRS (Hepatocyte growth

factor- regulated tyrosine kinase substrate) and NUMB (Protein numb homolog) in Human; and the epsins, EDE1 (EH domain-containing and endocytosis protein 1), ALY2 (Arrestin-like protein 2), AP180B (Clathrin coat assembly protein AP180B) and the actin cytoskeleton-regulatory complex proteins, PAN1 and SLA1 in yeast. Among yeast proteins of our dataset, a large fraction of residues is in LDRs, which increases the median disorder content of the ASP group (8.54% in yeast and 8.13% for Human). The median ratio of DBR residues also shows similar values for proteins of different functional groups in both species, indicating that the disordered regions of adaptor proteins are highly enriched in binding motifs (8.93% Human, 8.22% yeast).

Interestingly, the group of "enzymatic activity related proteins" (EARP) has the highest median disorder content among groups of the dataset of human proteins (22.84%) and the third highest median disorder content among groups of the dataset of yeast proteins (10.68%). The relatively high disorder content of EARP proteins might seem counterintuitive at the first, since enzymes are thought to be typically well-folded proteins. This is in fact the case for protein domains carrying enzymatic activity, such as the small GTPase enzymes. However, GTPase enzymes direct regulators, such as the long GAPs (GTPase activating proteins) are highly disordered (d.c. > 54%) in yeast (both ADP-ribosylation factor GTPase-activating proteins, GCS1 and GLO3) and considerably disordered in Human (d.c ranging from 22.03 to 52.22% for all four human GAPs). The two synaptojanins in Human, along with their yeast orthologs (Phosphatidylinositol 4,5-bisphosphate 5-phosphatases

INP51, INP52 and INP53) also show a large amount of structural disorder in regions outside their phosphatase domains. The domains responsible for the enzymatic activity occupy only a short segment of the long protein sequences (ranging from 946 to 1573 residues for the five proteins in the two species). The remaining disordered regions of human synaptojanins and their yeast orthologs are likely involved in protein-protein interactions. These proteins also have a high ratio of their residues in DBRs for Human (~15%). Synaptojanins in Human also show rather high LDR residue content (~28%), whereas for most yeast orthologs this value is around 18%. In Human, the EARP group includes also other (mainly clathrin pathway associated) enzymes that are largely disordered outside their enzymatic domains, such as AAK1 (AP2-associated protein kinase 1; d.c. 58%), auxilin (disorder content 45.24%) and GAK (Cyclin-G-associated kinase, disorder content. 39.51%).

The group of "neurotransmitter transport specific regulators" (NTSRs) – present only in Human – contains distinct protein families: synaptotagmins, complexins, several neurotransmission-specific SM proteins, synaptophysin and tomosyn. Complexins are the most disordered family of the entire membrane trafficking protein dataset; their disorder content ranges from 76.25% to 98.51%. However, the median disorder content for all the proteins in the NTSR group is only 8.65%, because except for the four complexins, all the other proteins are highly ordered.

Comparisons of the average disorder content in human proteins in the 7 functional categories to the corresponding functional groups of the yeast dataset showed lack of statistically significant differences in the average disorder content as assessed using the Wilcoxon Ran Sum Test. Additionally, the fraction of proteins with long disordered regions (LDR) in the different functional categories was compared among the species and against their corresponding proteome using Fisher's Exact Test. The comparison between species showed no significant differences at 5% significance level, with the exception of the OFRP group of human proteins with LDR ($\geq$ 100 residues), which was significantly larger than the corresponding yeast group. When comparing human proteins in different functional groups against the complete human proteome, the group of adaptor and sorting proteins (ASAP) was significantly enriched in proteins with LDR of the different lengths (LDR $\geq$ 30, 50 and 100 amino acids; p-values: 6.12E-03, 1.77E-02, and 9.73E-05, respectively). Similar results were observed for proteins related to enzymatic activity (EARP); a significantly high number of EARP proteins showed LDR of the three lengths (p-values: 4.95E-02, 3.28E-02, and 1.49E-03, respectively) with respect to the complete human proteome. The trend of enrichment in LDRs of the ASAP and EARP human groups was similar to the one reported for yeast proteins in these functional groups. A significant number of yeast adaptor and sorting proteins (ASAP) were enriched in LDRs of the three lengths (p-values: 3.13E-03, 4.79E-03 and 1.54E-02, respectively) with respect to the complete yeast proteome. The EARP group in yeast was also enriched in proteins with LDR of

length ≥ 30 amino acids (p-value = 1.42E-02) when compared to the complete yeast proteome.

4.4.5. Intrinsic disorder in the different vesicle-trafficking routes

We compared the disorder content of all budding and fusion-associated proteins in human and yeast involved in the three main vesicle trafficking systems regardless of their functional classification (Figure 4-3). Proteins associated to the clathrin-mediated route are the most disordered among the three systems, with 23.33% median disorder content in Human, and 22.58% median disorder content in yeast. When comparing proteins in the clathrin-mediated route to those in the COPI-mediated route, the difference in their average disorder content is significant for both species (Wilcoxon Rank Sum Test, p-value = 9.89E-03 in Human, p-value = 4.68E-02 in yeast.). Similarly, proteins involved in the clathrin-mediated route have higher average disorder content than those in the COPII-mediated route, and this difference is significant both in Human (Wilcoxon Rank Sum Test, p-value = 4.09E-02) and in yeast (p-value = 5.99E-03). The COPI and COPII-mediated routes show very similar disorder content in their corresponding proteins. The COPI-mediated route proteins have 9.20% and 8.22% median disorder content in Human and yeast, respectively, while in the COPII-mediated route the median disorder content in Human and yeast are 9.27% and 6.99%. Proteins in the COPI and the COPII-mediated routes do not present statistically significant differences in their average disorder content in either species.

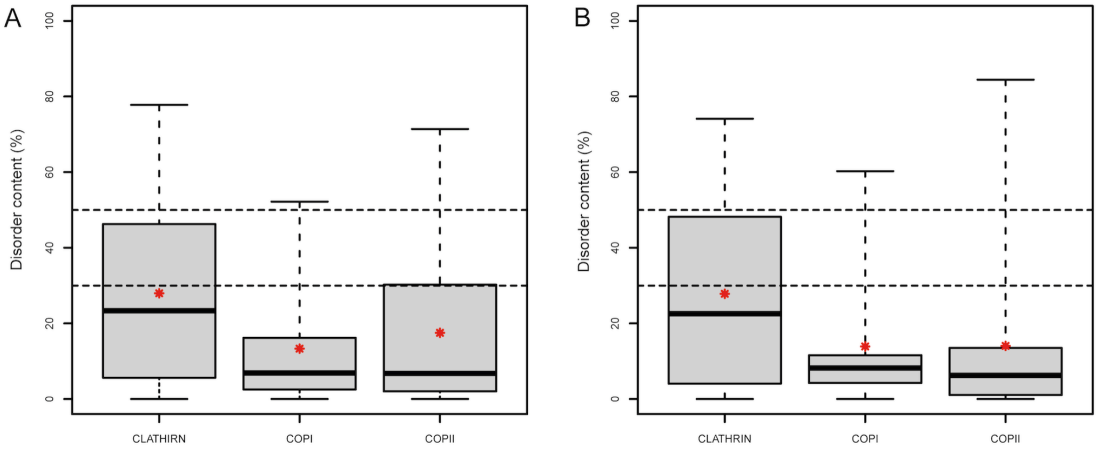

Figure 4-3. Disorder content for proteins involved in the three main vesicle trafficking pathways. Fraction (%) of predicted disordered residues (disorder content) calculated using IUPred for proteins involved in vesicle trafficking systems for Human (A) and yeast (B) for data reported in Table 4-1. The mean is depicted by a star. Proteins with disorder content (dc) (30%≤dc<50%) are considered fairly disordered; proteins with (dc≥50%) are considered highly disordered.

The mean and median values of all the disorder measures calculated for human and yeast proteins in the three vesicle trafficking pathways are reported in Table 4-1. Comparisons of the average disorder content in human proteins in each of the three pathways to the corresponding pathways of the yeast dataset showed lack of statistically significant differences in the average disorder content as assessed using the Wilcoxon Ran Sum Test. The higher disorder content of proteins in the clathrin-mediated route is partly found in the three clathrin light chains, which are highly disordered (disorder content ranging between 60.1% - 74.7%). In addition, the clathrin-mediated pathway is the only pathway containing several highly disordered proteins in ASP group, including the epsin protein family. In the

COPII-mediated pathway, most of the disorder content is found in outlier proteins: transport proteins Sec16A and Sec16B in Human, and transport protein Sec16 in yeast (disorder content > 54%). The COPI-mediated system also contains a few highly disordered proteins ArfGAPs (ADP-ribosylation factor GTPase-activating proteins; disorder content ~50%) in Human.

The ratio of proteins with long disordered regions for three pathways was compared to that of the complete proteome of each species. In Human, only the clathrin-mediated pathway showed a significant enrichment in proteins with LDR of all three lengths (LDR ≥ 30, 50 and 100 amino acids; p-values: 1.95E-04, 8.22E-04 and 7.77E-07, respectively). Similarly, the clathrin-mediated pathway in yeast was the only group enriched in proteins with LDR of the three lengths (p-values: 1.11E-03, 5.58E-04 and 2.91E-04, respectively).

4.4.6. Domains typically surrounded by disordered regions

We identified at least one Pfam entity (143 families; 153 domains; 3 motifs; 9 repeats) for 238 proteins in the human dataset. There were only 10 proteins for which no domain or family could be assigned. We further analyzed the Pfam patterns of highly disordered proteins, which are proteins with at least 70% disorder content or a high ratio of LDR residues (≥ mean+ 2StDev or ≥ 50% of disorder content). Examples include the complexins, which belong to the synaphin family and are highly disordered (disorder content 76-98%) and the family of

clathrin light chains, where both proteins have high predicted disorder content (60 and 74%) and have no folded domains assigned.

Our analysis also shows that there are several folded protein domains that are typically located in highly disordered proteins. These structured "islands" are usually surrounded by extended disordered regions at either or both sides, and are typically the only structural domain in the entire protein. Examples include the ENTH (epsin N-terminal homology) domain, the PID (phosphotyrosine interaction domain), the Sec16 domain, and the muHD (muniscin C-terminal mu homology domain). The three highly disordered epsin type clathrin adaptor proteins, for example, contain only an ENTH domain at the N-terminus of their sequence, while the remaining part of the protein is completely disordered. Epsins 1 and 3 also contain UIM motifs (Ubiquitin Interaction Motifs) and the adaptor protein- and clathrin-binding motifs within their disordered regions[279,296,274]. Another clathrin adaptor, DAB2 contains also one single domain at its N- terminus, the PID, which virtually corresponds to the only structured region of this highly disordered protein (d.c. 74%). Two other proteins in the dataset share this PID domain: NUMB (61.21%) and DAB1 (51.36%). PID spans the structured region of these proteins, while most of the remaining regions of the protein are completely disordered. The muHD domain is also coupled with a long disordered segment at the N-terminal side. The muHD domain is present in three disordered adaptors from the clathrin system: SGIP1 (SH3-containing GRB2-like protein 3-interacting protein 1, disorder content 62.68%), FCHO1 (FCH domain only protein 1, disorder content. 47.58%)

and FCHO2 (FCH domain only protein 2, disorder content 34.57%). Another example of these structured island domains is the Sec16 domain found in transport proteins Sec16A and Sec16. Sec16 domain is located approximately in the middle of these large proteins (~2000 residues), and it is surrounded by highly disordered terminal regions. The structural characteristics of Sec16A will be further discussed in next section. The ArfGAP domain is also usually located on the N-terminal end of the long, considerably disordered ArfGAP proteins (disorder content ranging between 52.22% and 36.63%).

There are other domains that are often surrounded by variable long disordered regions, but are also present in proteins that tend to have less disorder content. The BAR (Bin-Amphiphysin-Rvs) domain –involved in membrane curvature sensing– is present in amphiphysin (60.58% disorder content) and in several endophilins (A1, A2, A3, B1 and B2), which have substantially variable disorder content (7.12-35.60%). The protein kinase domain is present in the most disordered enzymatically active member of our human dataset, the AAK1 (AP2-associated protein kinase 1; disorder content 58%).

In summary, the vast majority of the protein domains that are always surrounded by highly disordered regions belong to the clathrin pathway ASP functional group of proteins. Their structural properties – mostly disordered with a single folded domain located at one of their termini – make them excellent candidates for the fly-casting mechanism for protein binding. Their long disordered regions have a bigger capture radius allowing them to efficiently span the

surrounding environment for their binding partners[96]. In fact, previous studies have shown that these adaptor proteins are able to form extended adaptor networks on the surface of the budding vesicle. Some of these adaptors are also involved in recruiting clathrin[297,279,296]. Additionally, the long disordered regions of these proteins contain a plethora of different binding motifs, which have been reported to facilitate specific interactions between the adaptor proteins, with clathrin, or with other components of the system[279,274].

To further investigate the role of disordered binding regions in building the adaptor network, we performed a systematic PDB search of clathrin coat specific protein complexes. We found several structures where the interaction of two clathrin adaptors is shown, and one of the partners uses its disordered binding regions to bind to the folded domain of the other partner. In the multisubunit AP-2, one of the most studied proteins involved in endocytosis, there are two long disordered regions connecting the two α-adaptin ear domains to the larger part of the complex. The recognition of cargo sorting signals is mediated by the larger part of the complex, while the principal clathrin-binding region is located in the disordered β2-adaptin hinge[279,274]. The two α-adaptin ear domains are favored targets of disordered tails of other clathrin adaptors and accessory proteins[314]. We found several distinct complexes forming these interactions (Figure4-4 A, B and C). In these complexes, usually very short peptides (with lengths ranging between 6-12 residues) of disordered regions in the partner proteins bind to the α-adaptin ear domain. In addition, we also identified a case where a relatively long, highly

disordered region of human protein stonin 2 binds to one of the small folded EF-hand domains of human protein EPS15 (Epidermal growth factor receptor substrate 15)(Figure 4-4 D).

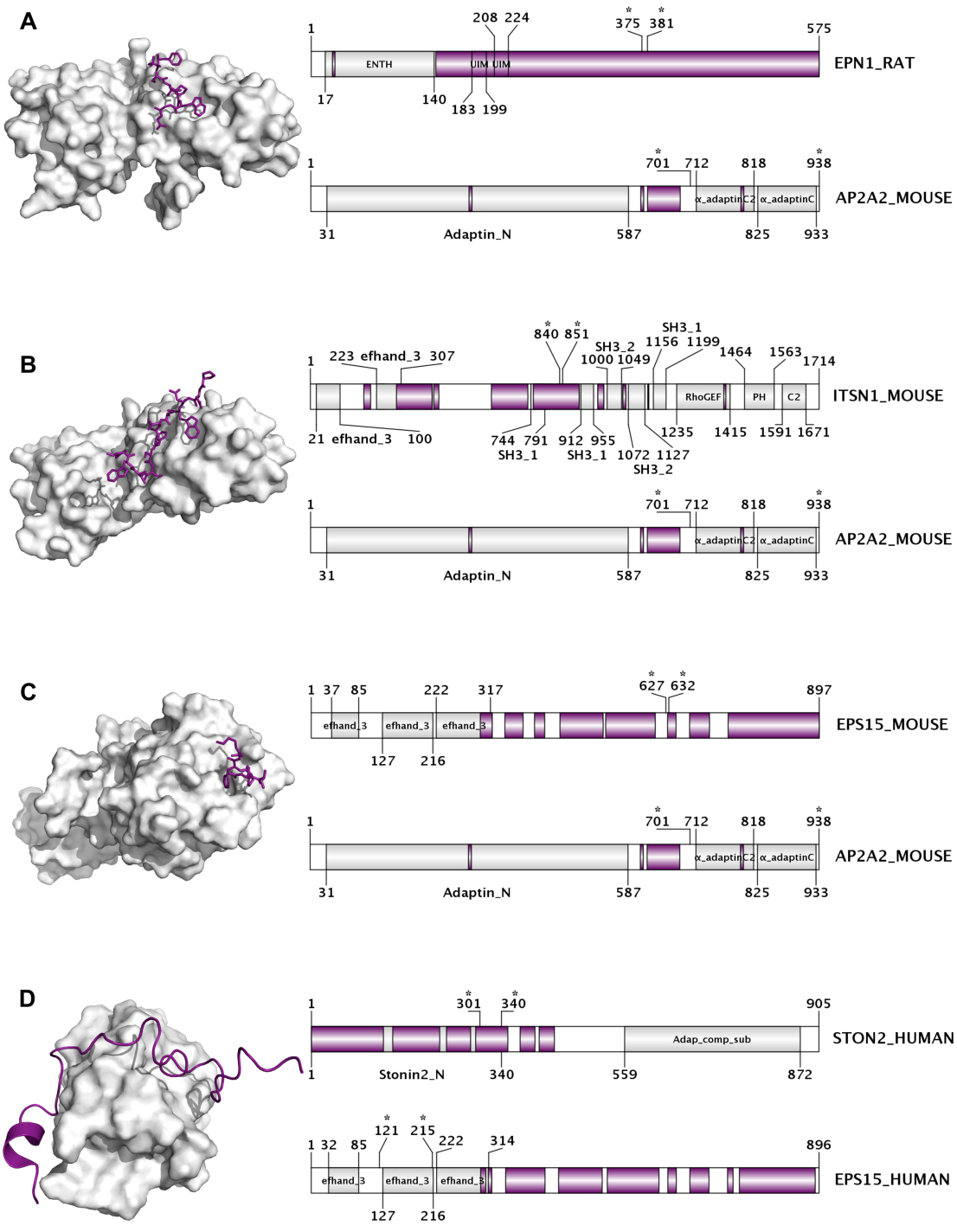
A
1
208 224
375 381
575
ENTH
UIM UIM
EPN1_RAT
17
140
183 199
1
701 712
818
938
α_adaptinC2
α_adaptinC
AP2A2_MOUSE
31
Adaptin_N
587
825
933
B
840 851
SH3_1
1156 1199
SH3_2
1000 1049
1464
1563
1714
1
223 efhand_3 307
RhoGEF
PH
C2
ITSN1_MOUSE
21 efhand_3 100
744 791
SH3_1
912 955
SH3_1
1072 1127
SH3_2
1235
1415
1591 1671
C
1 37 85
222
317
627 632
897
efhand_3
efhand_3
efhand_3
EPS15_MOUSE
127
216
D
301
340
905
Adap_comp_sub
STON2_HUMAN
Stonin2_N
559
872
121
215
1 32 85
222
314
896
EPS15_HUMAN

Figure 4-4. Interaction between clathrin-associated adaptor proteins. PDB reported complexes between two clathrin-associated adaptor proteins in which one of the adaptors interacts with a region predicted disordered in the unbound form. In the first three panels, the folded α2 subunit of mouse Ap-2 interacts with (A) rat epsin-1 (PDB 1KY6), (B) mouse intersectin-1 (PDB 3HS8) and (C) mouse EPS15 (Epidermal growth factor receptor substrate 15, PDB: 1KYF). In panel D, a relatively long disordered segment of human stonin-2 interacts with one folded EF-hand domain of human EPS15 (PDB: 2JXC). In each panel, the structure of the complex (left) and the domain maps for each interacting partner protein (right) are depicted. The top domain map represents the partner binding through the structurally disordered region. In panels A to C, disordered peptides are represented with sticks (purple) while the folded partner is shown in surface representation (white). In panel D, the long disordered segment of human stonin-2 is shown in cartoon representation. Domain maps for each interacting partner show names and locations of their reported Pfam domains. Disordered regions (length ≥ 3 residues) as predicted by IUPred are colored in magenta, while the structured segments are light-gray (predicted to be part of a Pfam domain) or white (if not predicted to be part of a Pfam domain). Regions present in the PDB structures are marked by stars.

In addition, we also found structures where other non-adaptor clathrin pathway associated proteins interact with AP-2 or clathrin via their disordered segments. Amphihpysin, for example, interacts with both (PDBs 2VJ0 and 1UTC) via two different disordered binding regions located in the long disordered segment following the BAR domain. Proteins from the EARP functional group also bind AP-2. In case of synaptojanin 1, two distinct constructs were shown to bind the α-adaptin domain (PDB ID: 1W80). The β subunit of AP-2 interacts with the PIP5K1C protein (PDB ID: 3H1Z).

4.4.7. Identification of orthologous protein pairs and analysis of their disorder content

We identified 56 human proteins that could be successfully matched to a yeast protein from our dataset. We focused on orthologous pairs involved in the budding and fission-associated functional groups because these groups show higher abundance of disordered regions. In addition, these functional groups involve the protein functions bearing the most distinguishable differences among the three pathways. We filtered the 56 proteins pairs choosing only those pairs in which at least one of the members showed considerably high disorder content (>30%). From the resulting 8 protein pairs, one showed very similar disorder content (less than 5% difference); 5 pairs showed more disorder in the human ortholog than in the yeast ortholog; and in 2 pairs the yeast protein showed higher disorder content than the human ortholog.

We analyzed the structural features of two protein pairs in detail (Figure 4-5) The first pair has the largest difference in disorder content among all the 8 pairs: human transport protein Sec24A and yeast protein SFB2 (SED5-binding protein 2), and they share 22.66% of sequence identity. In the other pair, both proteins are highly disordered: human transport protein Sec16A and yeast Sec16 (COPII coat assembly protein SEC16), and show 14.07% of sequence identity. According to the predicted disorder patterns of the Sec16 pair, their disordered pattern is well conserved (Figure 4-5 B). In the Sec24 pair (Figure 4-5 A), the human sequence is considerably longer. This difference in the proteins' lengths is due to a

protein segment in the N-terminal region of the human ortholog, which is missing in the yeast ortholog. This long, disordered N-terminal region in the human protein is also abundant in predicted disordered binding regions (shown in blue in Figure 4-5), and hence it might be considered a result of adaptive evolution. In fact, this subunit of the COPII coat-adaptor complex plays role in the recognition and binding of the cargo (transmembrane cargo proteins, and transmembrane cargo receptors of soluble proteins)[293]. Given that the repertoire of possible cargo proteins transported from the ER to the Golgi is considerably higher in human than in yeast, the presence of additional binding regions in the human ortholog is not surprising.

In the Sec16 ortholog pair (Figure 4-5B), both protein members are extremely long (>2000 residues), and highly disordered (74.44% and 71.4% disorder content in the yeast and human orthologs, respectively). The domain maps of these proteins show that the LDRs surrounding the Sec16 and Sec16_C Pfam domains are highly enriched in disordered binding regions. These two proteins are highly similar in length and can also be considered well conserved from a structural point of view: they both contain unstructured regions surrounding the conserved, structured domains. Their preserved disordered nature, with plenty of DBRs (54.8% in Human, 50.07% in yeast) and an even higher ratio of residues located in LDRs of at least 30 residues (62.1% in human and 68.2 in yeast), most certainly has an essential functional role. Sec16 is involved in the initiation of the COPII coat assembly and in the selection of cargo molecules. For the coat assembly, the LDRs can be especially advantageous, because – being able to bridge very long distances

through the fly-casting mechanism – they can reach for the components of the vesicle coat from the surrounding environment, and help the proper assembly of such components. In the clathrin-mediated system, the group of adaptor proteins I s usually responsible for this function. The adaptor proteins can utilize their disordered regions to form the adaptor network on the vesicle surface and to attach the clathrin chains to the surface of this network. In the COPII-mediated system, however, the adaptors are part of the multisubunit adaptor-coat complex, and the two subunits playing the adaptor role are not disordered enough to fulfill these roles. Thus, in the COPII-mediated system, a large disordered protein, such as Sec16 (and its homologs) is required to orchestrate the assembly of the coat components, especially when large distances need to be spanned.

Among the other 6 ortholog protein pairs, there are two pairs of synaptojanins, in which both human proteins are ~10% more disordered than their yeast orthologs. Another protein pair (human Sec24B; yeast Sec24) shows again more disorder in the human ortholog (~16% difference), although the difference is less striking that in case of the Sec24A-SFB2 pair. For the pair of clathrin- associated orthologous adaptors, human GGA3 (Golgi-localized, gamma ear-containing, ARF-binding protein 3) and yeast GGA2 (Golgi-localized, gamma ear-containing, ARF-binding protein 2), there disorder content difference is very moderate: 6% higher in the human ortholog compared to the yeast ortholog.

For the Sec31 COPII coat subunit ortholog pair (human Sec31A and yeast Sec31), the trend is inverted: the yeast protein is ~11% more disordered than the

human protein. In Human there are two paralog proteins sharing the same function, most likely distributing certain specialized tasks among each other. In yeast, Sec31 alone is in charge of all these functions. Thus, the higher disorder content in the yeast protein could have evolved in order to allow for multiple protein interactions via the moonlighting mechanism[74]. In the pair of GAP proteins consisting of human ArfGAP2 (ADP-ribosylation factor GTPase- activating protein 2, disorder content 47.2%) and yeast GLO3 (ADP-ribosylation factor GTPase-activating protein GLO3, disorder content 60.24%), the yeast ortholog was considerably more disordered than the human ortholog. These GAP proteins are large and highly disordered and act as GTPase activating factors of the ARF family small GTPases. They hydrolyze the Arf1-bound GTP, which may lead to the dissociation of the coatomer complex from the Golgi-derived vesicles, enabling the vesicle to fuse with the appropriate target membrane.

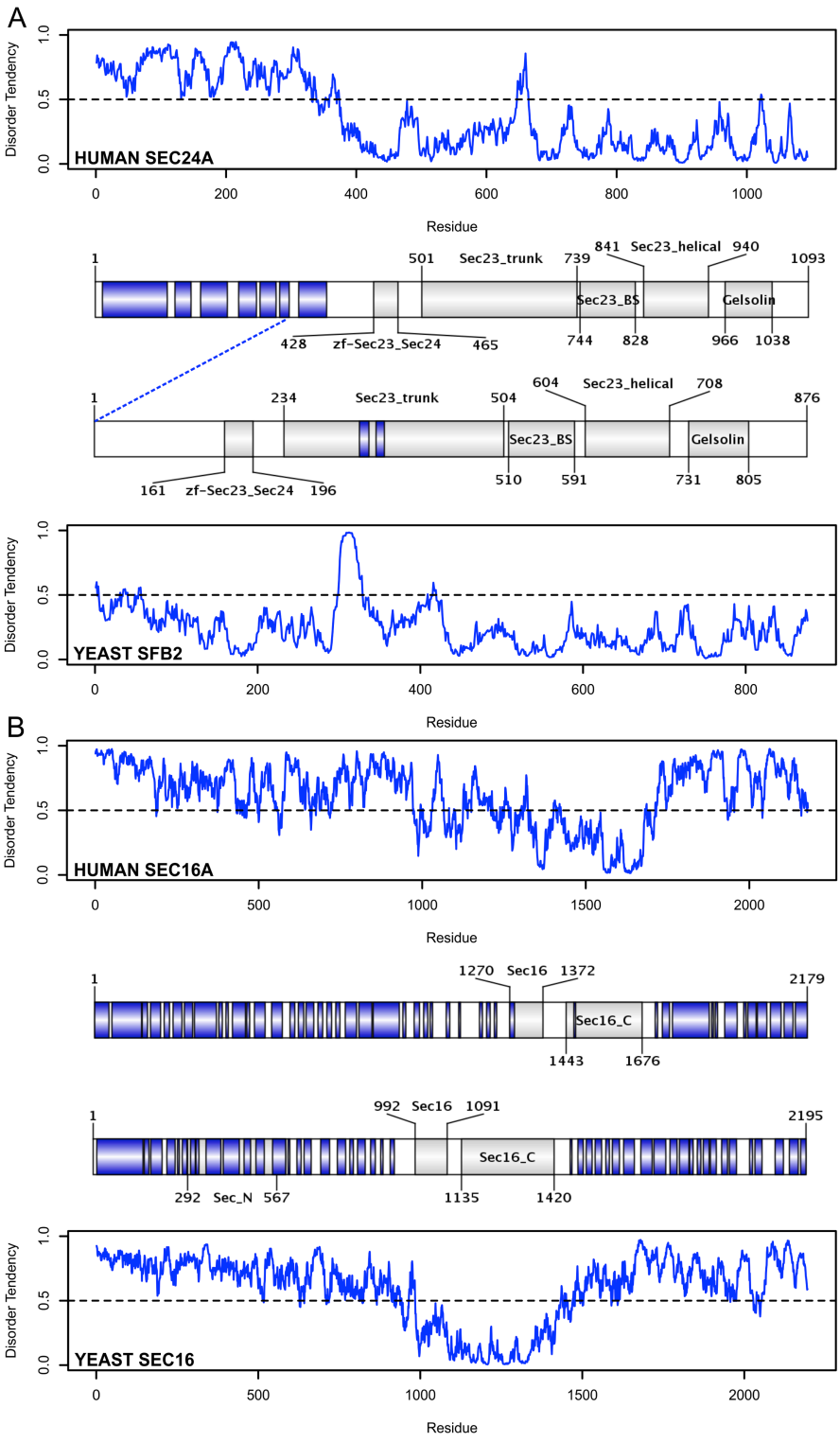
A
Disorder Tendency
1.0
0.5
0.0
HUMAN SEC24A
0
200
400
600
800
1000
Residue
1
501
Sec23_trunk
739
841
Sec23_helical
940
1093
Sec23_BS
Gelsolin
428
zf-Sec23_Sec24
465
744
828
966
1038
1
234
Sec23_trunk
504
604
Sec23_helical
708
876
Sec23_BS
Gelsolin
161
zf-Sec23_Sec24
196
510
591
731
805
YEAST SFB2
0
200
400
600
800
Residue
B
HUMAN SEC16A
0
500
1000
1500
2000
Residue
1
1270
Sec16
1372
2179
Sec16_C
1443
1676
1
992
Sec16
1091
2195
Sec16_C
292
Sec_N
567
1135
1420
YEAST SEC16
0
500
1000
1500
2000
Residue

Figure 4-5. Structural comparison of orthologous proteins involved in vesicle trafficking. Two protein pairs in the COPII vesicle trafficking system are presented. A) Moderately disordered (34.13% disorder content) human Sec24A COPII adaptor subunit and (disorer content 5.94%) yeast ortholog (SFB2, Sec24 related protein). B) Highly disordered human Sec16A (disorer content 71.41%) and yeast Sec16 (disorer content 74.44%) proteins. The predicted disorder (by IPred) is plotted (blue curve) with a order/disorder cut-off at y=0.5 (black dashed line). Residues with disorder tendency above this cut-off are considered disordered. A domain map of each protein shows the location and names of their identified Pfam domains (gray segments) and their predicted disordered binding regions (by Anchor) (blue segments). In each panel, the human ortholog is depicted in the top part; the disorder prediction curve followed by its corresponding domain map. The bottom part of each panel is a specular representation of the corresponding yeast ortholog: the disorder curve is topped by the domain map. Disorder curves and domain maps provide the structural information to define the disorder pattern. The blue dashed line connecting the domain maps in panel A shows the position in the human ortholog corresponding to the N-terminal end of its yeast ortholog.

4.5. Discussion

Vesicle trafficking routes have fundamental roles in eukaryotic cells, enabling transport of macromolecules between the various intracellular compartments and also between the cell and the extracellular environment. The COPI, COPII and clathrin-mediated vesicle trafficking routes comprise the major part of the cellular transport network, being responsible for the different types, locations and directions of traffic involved in endocytosis, the early and late secretory pathways and the retrograde Golgi-ER transport. Each trafficking route has its own well-conserved protein machinery and specific vesicle coat type with very few common proteins among them; however, they share several structural, mechanistic and

regulatory features. Despite these similarities, there are fundamental functional and evolutionary differences that strongly distinguish these routes, yet the molecular properties that could account for these differences had not been previously thoroughly described.

We provided a systematic assessment of the potential functional involvement of structurally disordered protein regions of proteins involved in the main vesicle trafficking systems in human and yeast. The results reported in Chapter 3 predicting vesicle-mediated transport proteins to be highly disordered in Arabidopsis, together with the functional requirements of proteins involved in vesicle trafficking, lead us to hypothesize that intrinsic disorder would be highly implicated in the vesicle trafficking systems of yeast and human cells. As discussed in the previous chapters, disordered regions have been widely recognized to be abundant in proteins related to signaling and regulatory roles[315,18]. Acting as flexible linkers between structured domains, disordered regions enhance the domains' free movement and rotations[316], providing the possibility for large, multidomain proteins to acquire multiple supertertiary structures[317]. Due to the increased accessibility of disordered regions and their enrichment in protein-protein interaction motifs[318] and posttranslational modification sites[300,90] in them, disordered regions are also frequently involved in molecular recognition and regulatory functions[279,274]. In addition, extended disordered regions are especially useful in the assembly of large macromolecular complexes[302], similar to the ones involved in vesicle trafficking.

The different measures of structural disorder used to describe the abundance and location of disordered protein regions in vesicle trafficking proteins allowed us to distinguish between major functional roles in which disordered regions could be involved. The overall disorder content of proteins provided a broad picture on the dependence of the proteins in different functional groups and trafficking route on structural disorder. The ratio of residues located in predicted DBRs offered an estimate of the involvement of disordered regions in protein-protein interactions. In cases where the ratio of residues in LDRs was considerably higher than the ratio of residues in DBRs, we could speculate that apart from promoting protein-protein interactions, disordered regions might also serve as flexible linkers between structured domains, or long spacers, assisting in the fly-casting mechanism by providing the possibility for the motif-rich parts to reach farther.

Despite the heterogeneity of proteins in the three major vesicle trafficking routes and the different functional groups, we still found that the human and yeast proteins followed similar overall tendencies of structural disorder in these functional groups and pathways. The systematic comparison of human and yeast proteins showed that proteins involved in vesicle trafficking tend to be more disordered in Human. However, the differences in most of the functional groups and pathways were statistically supported for a few cases (such as the OFRP functional group). The fact that human and yeast proteins belonging to the main vesicle trafficking pathways do not exhibit large differences in their average disorder

contents suggests that the process of vesicle trafficking exhibits essentially the same complexity in both organisms. Thus, in this case, the relationship of intrinsic disorder and complexity is more apparent at the biological process level than at organismal level thereby highlighting the role of disorder in proteins involved in vesicle trafficking.

The importance of disordered regions in proteins involved in vesicle trafficking is well reflected by the fact that almost all the main functional groups have highly disordered (>50%) proteins in both species. The large differences between the median disorder content of the proteins in the different groups nonetheless imply that certain functions require the presence of disordered protein segments more than other functions. Not surprisingly, most coat proteins are mainly ordered, since they tend to fold into rigid cage-like structures on the surface of all types of vesicles. However, some coat proteins were predicted to be largely disordered. The human clathrin light chains, for example, had the highest disorder content in the COAT functional group, followed by the different Sec31 COPII coat subunits. These highly flexible components of the coat may contribute greatly to an efficient coat assembly. In fact, structural flexibility has been observed to help the assembly of cytoskeleton, chromatin, and large protein complexes such as viral capsids[82,302]. The highly disordered nature of clathrin light chains, could also have an important role in the packing of the extraordinarily tight, highly overlapping, clathrin triskelion cage[276,319]. In addition, these disordered regions could be important for self- assembly[276].

Some of the proteins involved in fusion-related functions also showed a considerable amount of disorder, although most of the functional groups (MSTC, OFRP and NTSR) had rather low levels of median disorder content. The SNARE group was the most disordered among these in both species, reflecting the fact the different SNARE homology domains are unfolded in their monomeric form[280,281]. As previously discussed, many of the SNARE proteins, such as the syntaxin family members, also have disordered N-terminal regulatory segments that allow their regulatory binding partners (SM proteins) to modify their functions[281].

The NTSR group, although showing relatively low median disorder content, contains the most disordered protein family in the whole dataset: complexins. These SNARE regulatory proteins are predicted to be virtually completely unfolded. In their complexes, the central helix of complexins interacts with one SNARE complex, while the accessory helix forms a bridge to another SNARE complex, occupying the empty v-SNARE binding site to inhibit vesicle fusion. The accessory helix of complexins was recently shown to help organizing the t-SNAREs into a zigzag topology that is incompatible with fusion (PDB: 1KIL, 3RLO)[320,321]. Complexins prevent SNAREs from neurotransmitter release until an action potential arrives at the synapse. They are essential grappling/clamping proteins[322] that help stabilize SNAREs in an active, but frozen state, and only release SNAREs when synaptotagmins give a Ca^{2+}-induced signal[281,323]. The mechanism by which synaptotagmins can pass the information about the Ca^{2+} signal to complexins is not yet fully understood. According to our predictions, complexin regions forming the

helixes in the complexes are unfolded in their monomeric state, similar to the SNARE coiled coil homology domains they mimic.

The group of “adaptor and sorting proteins” showed the highest number of extremely disordered members, especially because of the presence of the non-complex-forming clathrin adaptor proteins. Although the ASP protein group contains many fully structured complex subunits as well (due to the many, highly similar subunits of the four different AP complexes in the clathrin route), it is evident that intrinsic disorder has a fundamental role in orchestrating many of the functions carried out by this group of proteins, such as linking the coat scaffold to the cargo and to the membrane, helping vesicle coat assembly by binding the coat subunits, and communicating with other accessory proteins. The dependency on structurally disordered regions of the ASP group is especially high in case of the clathrin system, since most of the individual clathrin adaptors, many of the accessory proteins and also some of its enzymes (like synaptojanins) have extremely long disordered tails with many protein interaction motifs. In addition, when analyzing those solitary folded domains, which behave like structured “islands” and are surrounded by extended disordered regions on either or both sides, most of the examples we identified belonged to the ASP group of the clathrin-mediated system. The proteins with long disordered segments, enriched in molecular recognition features are good candidates for the fly-casting mechanism. The proteins could behave as a fishing stick, their folded domain being fixed to the surface of the vesicle or to bigger protein complexes, while their disordered, flexible

tail could freely reach for their various protein partners over relatively long distances. Proteins with long disordered regions have a large capture radius that can help them efficiently utilize their many interaction motifs. This binding mode can be especially advantageous in the vesicle assembly process because it may enhance the speed of recognition due to the larger capture radius and may bring the coat components into close proximity to the surface of the budding vesicle. The interaction specificity provided by these interaction motifs has been reported to be essential in the assembly of other macromolecular complexes[7].

Finally, we collected several protein complexes from the PDB that provide structural evidence for protein interactions mediated by the induced folding of disordered binding regions in clathrin system related proteins. Many of these structures showed the same protein domain, the AP-2 α-adaptin ear domain, facilitating specific interactions with disordered binding motifs of its partner proteins. The partners not only included adaptor proteins; there were also structures of synaptojanin-1 and amphiphysin interacting with the ear domain. We identified other examples of clathrin system related complexes too, such as the human stonin 2 binding to the EPS15 EF- hand domain. All these observations are in agreement with previous results describing extended, dynamic protein network on the surface of clathrin coated pits[297]. The composition of this protein network is probably highly variable in a localization-, route-, and even cargo-specific manner, with several different functional groups among its members.

The functional importance of disordered regions in vesicle trafficking proteins was also well reflected by the conserved nature of the location of these regions, while the variability in their length and their low sequence similarity accounted for their increased adaptability and tolerance against mutations compared to folded protein domains. When investigating ortholog protein pairs in human and yeast, the location of disordered regions was found to be quite conserved, while the length of the disordered regions appeared to be more subject to evolutionary change. In case of the Sec16 pair, the long disordered regions surrounding the structured domains were very well preserved, and even the lengths of the two proteins were highly similar. Since almost the full length of the two long disordered "arms" of the proteins was enriched in predicted disordered binding sites in a well-conserved way, they are likely essential in the initiation of the COPII coat assembly. The level of conservation in these protein regions seems to largely depend on the specific functional needs of the given protein. In case of the Sec24 orthologous pair, for example, the human sequence had a considerably long N-terminal unstructured region, which was virtually missing in the yeast ortholog. The presence of numerous predicted disordered binding regions in the human ortholog suggests that this region is the result of adaptive evolution. Since this protein is a key player in cargo recognition and binding –which obviously involves a larger repertoire of possible cargos in human– the emergence of such adaptive regions are indisputably beneficial.

Results showing that proteins in the clathrin system are significantly more disordered than proteins in the COPI and COPII systems not only imply the larger dependence of the clathrin system on disordered protein segments and support the concept of highly dynamic networks formed by its proteins, but also accounts for the differences between the three routes from the evolutionary point of view[296]. Disordered regions not only provide conformational freedom but also a kind of evolutionary freedom. The increased tolerance against mutations gives disordered regions the possibility of fast evolutionary changes, thus providing exceptional adaptability. As reported, the clathrin-mediated system shows marked plasticity and robustness compared to the COPI and COPII systems. There are many observations emphasizing the increased adaptability of the clathrin-mediated route. It exhibits many species-specific characteristics[288,292], and it has been extensively modified to assist other specialized pathways. Adaptor proteins and clathrin itself, for instance, are often manipulated to create novel types of organelles, including the rhoptry secretory organelle in *Toxoplasma gondii*[324], the contractile vacuoles of *Dictyostelium* species[325], special vesicles for odorant receptors transport of *Caenorhabditis elegans*[326], and the machinery for sorting proteins to the basolateral plasma membrane of vertebrate epithelial cells[327]. Biogenesis of synaptic vesicles in animals and in human is also performed by endocytic adaptors[328]. Similarly, there are other organelles that also require these adaptor proteins for their maturation[329]. Apart from the species and tissue-specific inventions, clathrin system adaptors are also frequently used for various functions during embryonic development[329,330].

Taken together, these observations on the many different adaptive changes of clathrin-route related proteins strongly support the idea that this route more versatile than the COPI and COPII systems. The structural background of this adaptability had not been explored until now; here we suggest that the elevated level of structural disorder found in proteins from the clathrin-mediated route provides a good explanation for the exceptional adaptability of this pathway. Furthermore, most likely intrinsic disorder can be accounted for the high evolutionary plasticity of clathrin-associated proteins.

We have identified the implications of protein disorder in the different protein groups involved in the main vesicle trafficking routes in human and yeast. However, we need to take into account that the classification schema dividing proteins according to their functional roles in the major membrane trafficking routes was in a way artificially designed, and may not always reflect the real situation in the cell in an accurate manner. Potentially misclassified protein sequences might account, for example, for the large deviations in disorder content shown in some of the protein groups. In addition, some functional groups exhibited very dissimilar number of proteins, which might also introduce some bias in their comparison.

We found many functional modalities enabled by disordered regions present in vesicle trafficking proteins. These include regulatory roles, the use of flexible linkers, mediating protein-protein interactions, and the quick assembly of large macromolecular complexes by fly-casting. Taken together, our results provide

compelling evidence of the functional involvement of structural disorder in the proteins from the main vesicle trafficking systems. The presence of highly disordered proteins in almost all the main functional groups of vesicle trafficking proteins emphasized the unquestionable importance of disorder for this cellular process. In addition, the differences in intrinsic disorder abundance between the proteins in the three main trafficking routes provided the structural background for long standing observations on the functional and evolutionary differences of these vesicle trafficking systems.

Chapter 5

Intrinsic disorder and protein packing defects as promoters of protein interactions

The following chapter is based on the article Published in: Natalia Pietrosemoli; Alejandro Crespo; Ariel Fernández; *J. Proteome Res.* **2007,** 6, 3519-3526, DOI: 10.1021/pr070208k.

5.1. Introduction

In Chapter 3 we analyzed the role of intrinsic disorder at a genomic scale, while Chapter 4 focused on the role of intrinsic disorder in a specific cellular process. In this chapter, we will discuss unstructured regions at protein level. This seminal work first lead us to question the extent of the implication of unstructured regions in protein function and in protein-protein interactions. Later, the role of

disordered regions in promoting interactions became of major interest in the disorder field[74,182,331].

In this work, we proposed that weaknesses in the protein's backbone hydration shell signaled structurally unstable regions, and that such regions promoted protein associations and intermolecular interactions to become more stable. We assessed the vulnerability of protein backbone by calculating their dehydrons. Dehydrons had been recently defined as water-exposed intramolecular backbone hydrogen bonds that are not "wrapped" by a sufficient number of nonpolar groups[332,333]. They represent structural singularities or "packing defects" [334,335,336] of proteins, since structured regions of proteins need to exclude water from their amide-carbonyl hydrogen bonds in order to maintain their fold[337,334]. Studies on these dehydrons showed that they favored the removal of surrounding water in order to strengthen and stabilize the underlying electrostatic interaction, and thus were thought to be implicated in protein associations and macromolecular recognition[334,335,336,332,338,339,340,341,342]. Insufficiently wrapped intramolecular hydrogen bonds became stronger and more stable by the attachment of a ligand or binding partner that could further contribute to their dehydration.[333,340]. Moreover, it was observed that dehydrons were key players for driving association: they were crucial in a good portion of the PDB complexes reported at the time (~38%) and still of significant importance in about 95% of the complexes[334,336].

In this work, we scanned the PDB for proteins with large clusters of dehydrons, and we reported on the functional role of these regions. The presence of

these large concentrations of packing defects in a soluble protein signaled structural singularities (as strong dielectric modulators, i.e., quenchers of the local dielectric permittivity) characterized as intermediates between ordered and disordered regions. The potential functional implications of these singularities were investigated. While order and intrinsic disorder were already well characterized structural attributes of protein sequences[70], we proposed that these unstable regions of soluble proteins represented a novel category. According to our results, these protein regions have distinct properties compared to both globular proteins and disordered regions in general, because they could not form enough favorable intrachain interactions to fold on their own, and were likely to gain stabilizing energy by forming binding partnerships[107,343]. Our observation that such protein regions in the twilight of order and disorder fostered associations, was very related to later observations on disordered protein regions that became order upon binding to a partner[8,97]. We reported that these unstable protein regions have distinct novel properties compared to both globular proteins and disordered regions in general, because they could not form enough favorable intrachain interactions to fold on their own, and were likely to gain stabilizing energy by interacting with a globular protein partner[107,343].

5.2. Hypothesis

Large clusters of packing defects in soluble proteins constitute structural singularities that are intermediate between ordered and disordered structures and act as promoters of protein interactions.

5.3. Methods

The general workflow of the analysis consisted in collecting the non-redundant domains reported in the PDB, calculating their packing defects (i.e the amount of wrapping of their backbone hydrogen bonds), and predicting their disorder propensity. Then, we calculated the "wetting parameter", i.e., the number of hydrogen-bond partnerships involving water molecules hydrating each domain and correlated it to their wrapping.

5.3.1. Dataset

The dataset for the analysis was constructed extracting single-domain proteins from the Protein Data Bank[309] (PDB). This dataset was filtered out for redundant and homologous domains, resulting in 2982 domains with less than 50% identity in aligned sequences.

5.3.2. Protein disorder predictions

Disorder predictions were based on the PONDR-VLXT[60] program. This method (Section 1.2.4), assigns a disorder probability ($0 \leq f_d \leq 1$) to each residue

within a sliding window (of 40 residues), representing the predicted propensity of the residue to be in a disordered region (f_d = 1, disorder; f_d = 0, order).

5.3.3. Identification of packing defects in soluble proteins

The extent of wrapping of the hydrogen-bonds, ρ, was quantified by determining the number of non polar groups contained within a desolvation domain typically defined as two intersecting balls of fixed radius (thickness of three water layers) centered at the α-carbons of the residues paired by the amide-carbonyl hydrogen bond. A cluster of packing defects was defined as the maximal set of dehydrons with intersecting desolvation domains.

5.3.4. Hydrogen-bonding partnerships for interfacial water

We calculated the thermal average <Γ>, of the average number of hydrogen-bond partnerships involving water molecules for the p53 transcription factor DNA-binding domain. This domain was selected because it contains three of the largest dehydron clusters to be found in PDB. The calculations on the thermal average were obtained from the trajectories generated by 5 ns molecular dynamics (MD) simulations performed with the GROMACS 20 program[344]. The Γ-values were determined for each water molecule within a 6 Å radius spherical domain centered at the α-carbon of each residue. The starting geometry was adopted from the monomeric p53 structure 2GEQ from PDB. A similar strategy was followed to include the three additional representative domain folds, thus covering the major topology classes: the SH3 domain (all β-stranded; PDB 1SRL); ubiquitin (α/β; PDB

1UBI), and λ-repressor (all α-helical, PDB 1LMB). For all given cases, the staring conformation was embedded in a pre-equilibrated cell of explicitly represented water molecules and counterions[345,346]. Then, the entire system was equilibrated for 5 ns. Computations were performed by integration of Newton's equations of motion with time step 2 using GROMACS in the NPT ensemble with box size 8 × 8 × 8 nm^3 and periodic boundary conditions. The box size was calibrated so that the solvation shell extended at least 12 Å from the protein surface at all times. The long-range electrostatics were treated using the Particle Mesh Ewald (PME) summation method.21 A Nose-Hoover thermostat was used to maintain the temperature at 300 K, and a Tip3P water model with OPLS (Optimized Potential for Liquid Simulations) force field[345,346] was adopted.

5.4. Results

5.4.1. Insufficiently wrapped intramolecular hydrogen bonds are associated to twilight regions of intrinsic disorder

According to our calculations, structures of soluble proteins have at least two-thirds of their backbone hydrogen bonds wrapped on average by ρ = 26.6 ± 7.5 nonpolar groups for a desolvation sphere of radius 6 Å. We defined our dehydrons as those hydrogen bonds contained in the tail of the distribution (i.e., the mean, ρ = 26.6 minus one standard deviation σ = 7.5[333]), hence those with a microenvironment of 19 or fewer non polar groups.

There is a strong correlation between the extent of wrapping of intramolecular hydrogen bonds (ρ) engaging a given residue (if any), and its intrinsic disorder propensity (f_d) (Figure 5-1). Such correlation suggests that clusters of dehydrons correspond to protein regions lacking structural integrity.

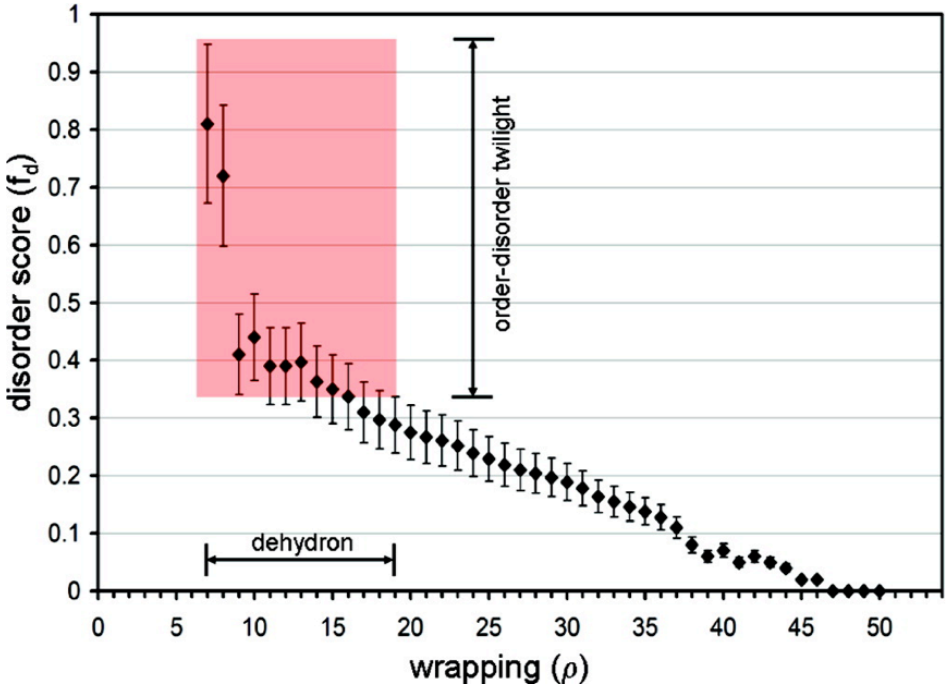

Figure 5-1. Correlation between intrinsic disorder of a residue and the extent of wrapping (ρ) of the backbone hydrogen bond engaging that particular residue (if any). Intrinsic disorder was predicted for each individual residue of 2982 non homologous PDB domains. Residues were independently grouped in 45 bins, according to the extent of wrapping ($7 \leq \rho \leq 52$). The average score has been determined for each bin (square), and the error bars represent the dispersion of disorder scores within each bin. The strong correlation between the disorder score and the extent of wrapping and the dispersions obtained implies that dehydrons can be safely inferred in regions where the disorder score is $f_d > 0.35$. The red rectangle represents the order–disorder intermediate region where the existence of dehydrons ($7 \leq \rho \leq 19$, for desolvation radius 6 Å) may be inferred from the disorder score. No hydrogen bond in monomeric domains reported in PDB was found to have less than 7 wrappers, implying a threshold for structural sustainability in soluble proteins. Figure from [21].

5.4.2. Clusters of packing defects and discrete solvent effects

Insufficiently wrapped backbone hydrogen bonds in a soluble protein are partially exposed to solvent, and they have been observed to favor the removal of hydrating molecules in order to enhance the polar-pair electrostatics[334,332]. Moreover, the resulting bond stabilization overcomes the amount of work needed to remove the solvating water molecules[335,332]. In order to describe this dehydration propensity, we calculated the extent of constraint of the interfacial water molecules, Γ. This parameter corresponds to the average number of hydrogen-bonds partnerships involving water molecules that are within the desolvation domain of each residue in the protein structure ($0 \leq \Gamma \leq 4$). Results of the calculations of the thermal average $\langle\Gamma\rangle$ for each residue belonging to the DNA-binding domain of p53, where if no water was found in the disolvation domain, the bulk water value ($\Gamma = 4$) was adopted are shown in Figure 5-2. Three dehydration hot spots are observed: hot spot 1 (residues 171-181), hot spot 2 (residues 236-246), and hot spot 3 (residues 270-289). In parallel, the structural representation of the backbone (depicted as virtual bonds joining the alpha carbons of the residues, in blue) of the p53 domain, along with its dehydron pattern (depicted as virtual bonds joining the alpha carbons of residues paired by backbone hydrogen bonds, in green) is shown in Figure 5-3. A comparison of Figure 5-2 and Figure 5-3 shows hot that the location of the three major clusters of dehydrons of the p53 domain corresponds to its dehydration spots. Moreover, these dehydration hot spots coincide with residues in the dimerization surface such as the Arg 178 in each monomer (present in hot

spot 1), and in DNA recognition via arginines in position 245 (present in hot spot 2), 270 and 277[347] (present in hot spot 3).

Dehydrons were also computed to obtain representatives of the additional domain folds: SH3 domain (2 dehydrons, PDB 1SRL), λ-repressor (26 dehydrons, PDB 1LMB) and ubiquitin (16 dehydrons, PDB 1UBI). The dehydron patterns obtained for these domain folds were consistent with the results obtained for p53's DNA-binding domain, thus we proposed that dehydrons emerged as the dehydration hot spots on the protein interface.

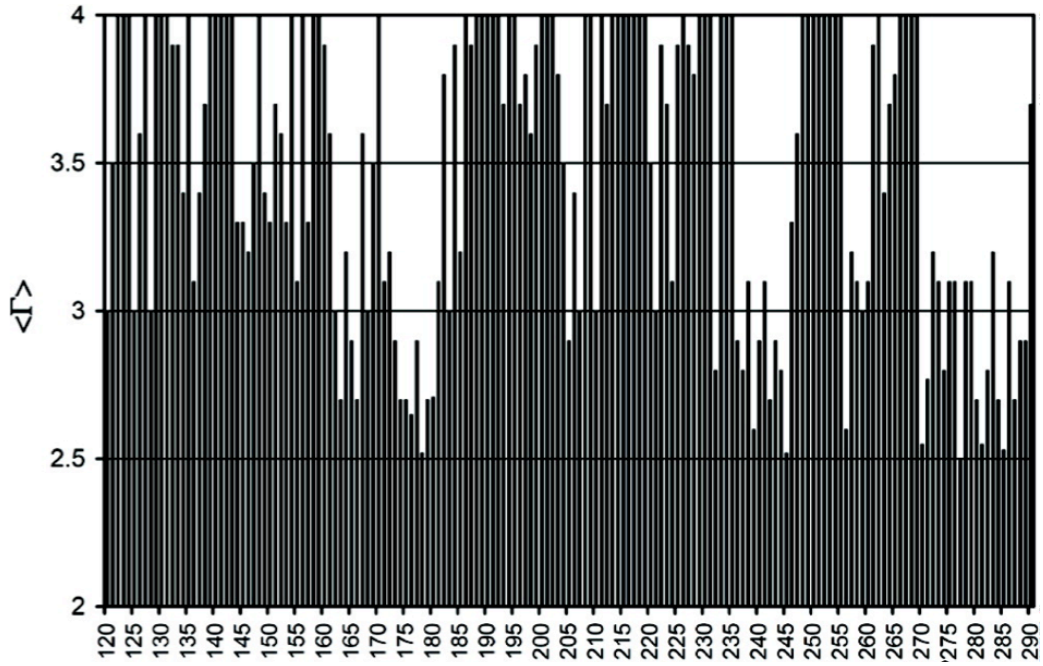

Figure 5-2. Thermal average of the average number of hydrogen-bond partnerships, <Γ> for water molecules within the desolvation domain of each residue in the DNA-binding domain of p53. If no water is found in the desolvation domain (i.e. buried residue), the bulk water value Γ = 4 is adopted. Figure from [21].

To study the solvating-water confinement induced by a packing defect, we selected a water molecule within the desolvation domain of Arg277, which is paired with a dehydron to Arg 280 (Figure 5-3) and performed 1 ns molecular dynamics simulations equilibrating the protein chain with surrounding water (as described in section 5.3.4). A snapshot of this water molecule in Figure 5-4 shows that it has three hydrogen-bond partners: two with neighboring water and one with the Arg 277 backbone carbonyl. We selected a 3.6 Å as the threshold for the hydrogen distance between heavy atoms. Because of the incomplete wrapping of dehydron ARg277-Arg280, the nearest water molecule is found at 2.8 Å between carbonyl and water oxygen atoms. While the water molecule is engaged with the Arg277 backbone carbonyl, it is deprived of one hydrogen bond partnership when compared with bulk water (Figure 5-3 and Figure 5-4).

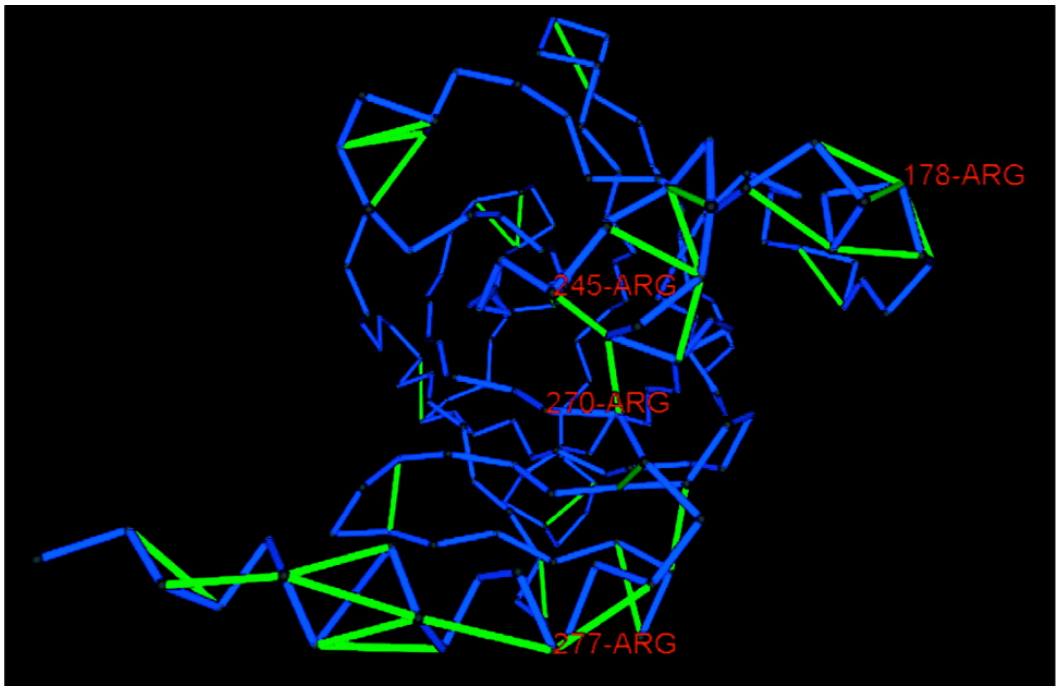

Figure 5-3. Dehydrons for p53 DNA-binding domain. The backbone is indicated by blue virtual bonds joining α-carbons and dehydrons are shown as green segments joining the α-carbons of residues paired by backbone hydrogen bonds. Figure from [21].

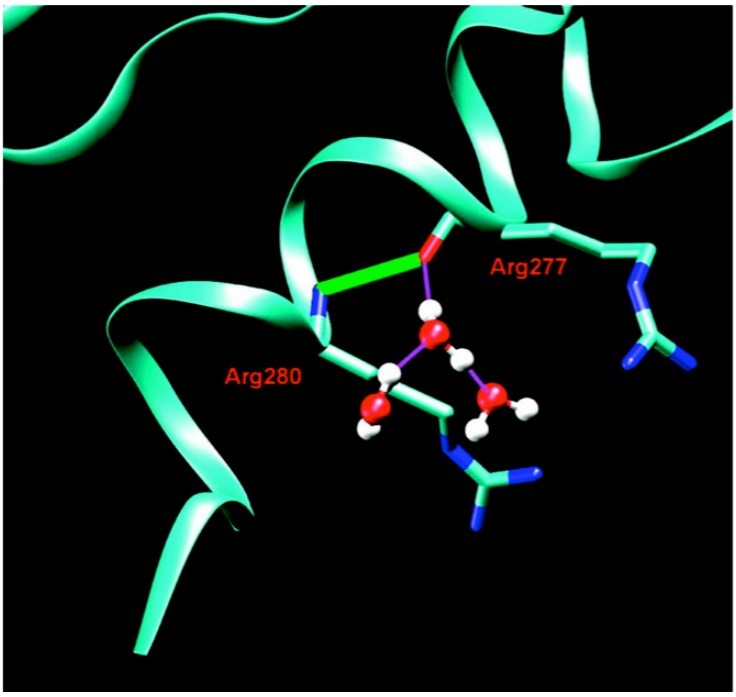

Figure 5-4. Snapshot (after 1 ns of MD) of a solvating water molecule and its hydrogen bond partnerships (purple bonds) within the desolvation domain of Arg277 in the DNA-binding domain transcription factor p53 (ribbon representation, fragment). The backbone amide−carbonyl dehydron Arg277−Arg280 is shown in green. Figure from [21].

In order to determine a generic relation between ρ and Γ, we also analyzed the three additional single-domain folds representing the main protein topologies: SH3-domain; ubiquitin and λ-repressor. The correlation between wrapping and dehydration propensity (Figure 5-5), has the following characteristics: i) dehydrons ($\rho \leq 19$) generate Γ-values in the range $2 \leq \Gamma \leq 3.6$; ii) no PDB hydrogen bond is wrapped by $\rho < 7$; iii) the upper wrapping bound, $\rho = 28$ corresponds to bulk-like water ($\Gamma = 4$) in the desolvation domain; and iv) all solvating water is excluded from the desolvation domain for $\rho > 28$.

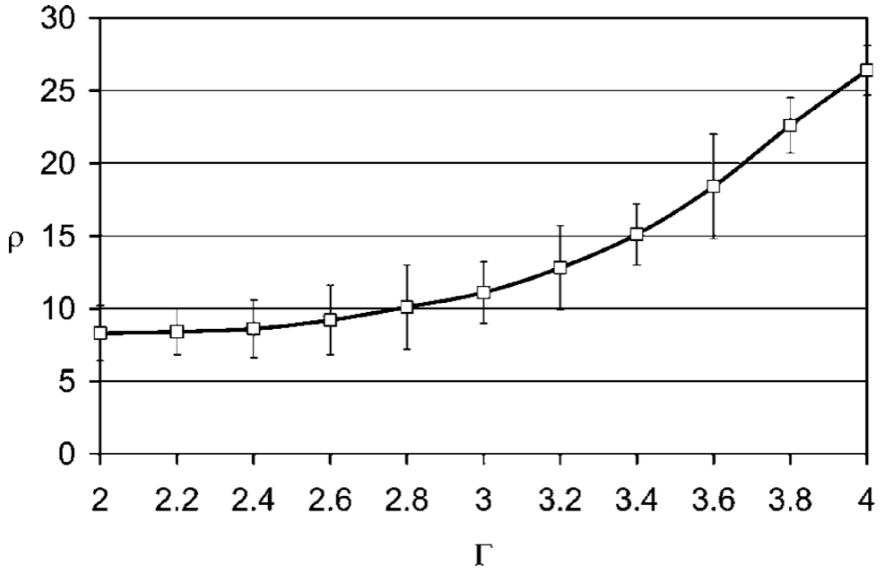

Figure 5-5. Correlation between hydrogen-bond wrapping ρ and wetting parameter Γ. Each residue is assigned a ρ-value averaged over all backbone hydrogen bonds in which it is engaged. Data extracted from the wetting computation on the p53 DNA-binding domain and three additional folds: the SH3 domain (2 dehydrons, PDB 1SRL); ubiquitin (16 dehydrons, PDB 1UBI), and λ-repressor (26 dehydrons, PDB1LMB). Figure from [21].

5.4.3. Defective packing and dielectric modulation

According to our analysis, the dielectric modulation is promoted by discrete solvent effects derived from the insufficient packing of the protein backbone. These discrete effects cannot be captured properly by conventional continuous models, which should be adapted to deal with local dielectric modulations. This dielectric modulation refers to the quenching in the local dielectric permittivity, and it is caused by the local reduction of the hydrogen-bond partnerships of solvating water molecules[348]. We quantified the analytic dependence of the dielectric permittivity on the wetting parameter Γ: $\varepsilon = 1 + \chi(\Gamma)$ (for further details, see [21]). The dielectric

quenching is extreme upon moderately small losses in hydrogen-bond partnerships (Figure 5-6). Accordingly, the most dramatic decrease the curve is marked by a drop in ε-values from 50 to 7 as Γ is reduced from 3.5 to 2.5.

Results shown in Figures 5 and 6, allowed us to conclude that clusters of packing defects (where $\rho \leq 19$) serve as potent enhancers of the electric fields generated at the protein interface. Accordingly, the typical loss in hydrogen-bonding partnerships associated with dehydrons solvation places Γ in a range of $2 \leq \Gamma \leq 3.6$. This interval contains the region of most dramatic dielectric quenching, decreasing the permittivity by an order of magnitude with respect to bulk water. In turn, this effect translates in an order of magnitude increase in electrostatic interactions.

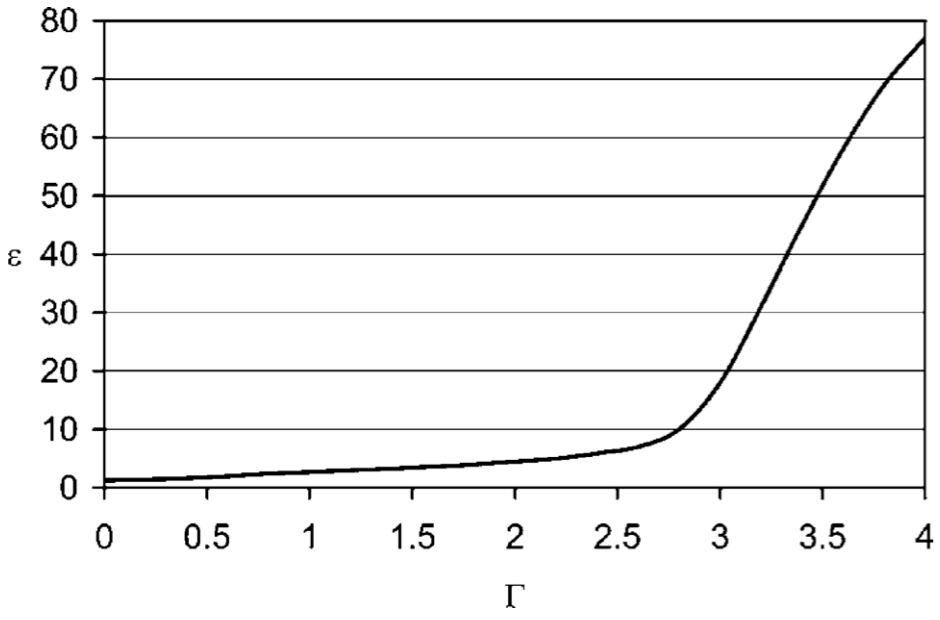

Figure 5-6. Analytical dependence of the dielectric permittivity ε on the wetting parameter Γ. Figure from [21].

5.4.4. Discrete dielectric quenching in the p53 DNA-binding domain: a study case

We further explored the functional significance of the three dielectric modulators in the DNA-binding domain of p53 by examining its dimeric state and its role as transcription factor. Figures 5-3 and 5-7 show that there is a cluster involving 5 dehydrons (173-176,174-178, 175-178, 176-179, and 178-180), which is actually located at the dimer interface. Due to their dehydration propensity and their role as promoters of protein associations[334,332,340,341,342] this clustering of packing defects fosters p53's dimerization. The dimerization involves a resonant pairing of the Arg178 from each monomer (Figure 5-7) likely to promote supramolecular charge delocalization with distal charge separation at all times. Significantly, <Γ> reaches a minimum precisely at Arg178 (Figure 5-2). Other additional minima in <Γ> correspond to Arg245, Arg270, and Arg277 (Figure 5-2). These arginines have a crucial role in DNA recognition[347]. Arg245 is engaged in the dielectric quenching region 236−246, while it is also a part of one of the dehydron clusters (236-239, 237-245, 237-271, 239-242, 239-244, and 240-242) shown in Figure 5-3. Similarly, Arg270 and Arg277 belong to dielectric quenching region 270−289 and they involved in one of the largest dehydron clusters found in the PDB (237-271, 274-277, 277- 281, 280-284, 281-285, 282-285, and 285-288 (Figures 5-3 and 5-8).

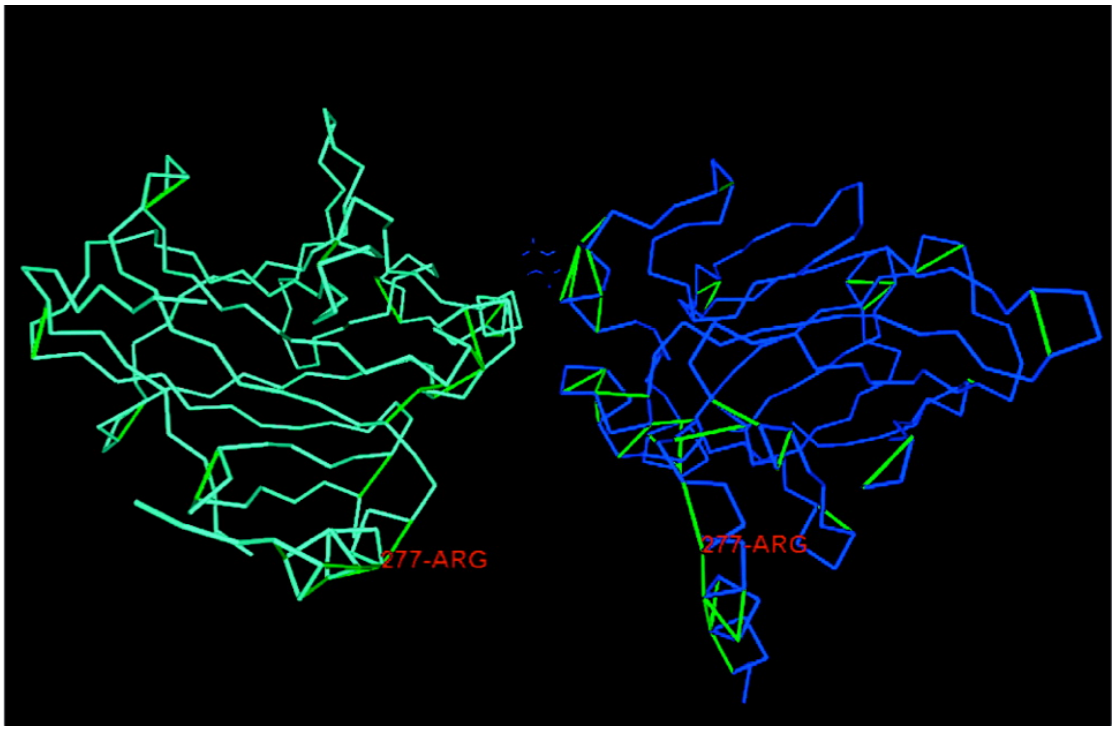

Figure 5-7. Backbone and dehydron representation of the dimmer interface for the DNA-binding domain of p53 (PDB 2GEQ). The side chains of the Arg178 of each monomer involved in a resonance pair are shown. Figure from [21].

While examining the protein–DNA complex of the DNA-binding domain of p53 (PDB 2GEQ), it is observed that the three residues directly involved in DNA are a arginines in positions 245, 270 and 277(Figure 5-8), in accord with previous works[347]. The two latter interact with the negatively charged backbone phosphates, while Arg 277 acts as intra-base "intercalator". These observations show that the electrostatics of protein-DNA recognition not only imply matching charges along the geometrically compatible interfaces, but also require a device to promote dehydration at the protein-nucleic acid interface. This enhancing of the electrostatic

recognition is indeed achieved through the large dehydron clusters surrounding the three arginines that are directly implicated in the protein-DNA association (Figures 5-3 and 5-8). Thus, the fact that the three arginines involved in DNA recognition are also dehydration hot spots is not adventitious, but a functional requirement for the transcription factor.

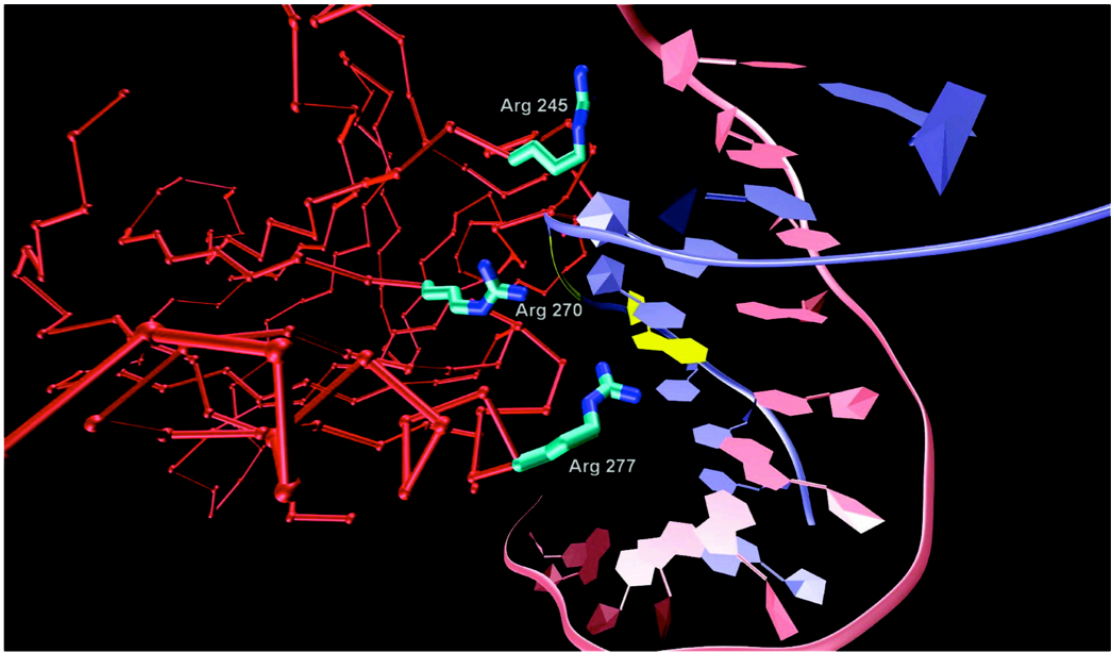

Figure 5-8. Protein–DNA complex of the DNA-binding domain of p53 (PDB 2GEQ). Side chains of the key residues directly implicated in DNA recognition, Arg245, Arg270, and Arg277 are shown. The pyridine base recognized by Arg277 is shown in yellow, whereas the individual DNA strands are shown in lilac and light magenta.

5.4.5. The most defectively packed protein domains

The non-redundant, non-homologous dataset of proteins was grouped according to the size of the dehydron cluster, *n* in each structure. The groups intersect to a considerable extent (Figure 5-9 (inset), and their respective populations decrease monotonically with an approximate power law $n^{-1/2}$. Each n-group was divided into five non-disjoint functional categories: biosynthesis, enzymology, cell signaling, cytoskeleton, and cancer. We normalized the contribution of each category to each n-group by taking into account the relative abundance of each category in our curated database. Thus, the number of PDB-domains within an n-cluster in a functional category was divided by the total number of PDB domains in the category. The relative abundance (%) of each functional category for each n-group is reported in Figure 5-9. The distribution of dehydron clusters becomes a marker of the functional categories, where biosynthesis peaks at 2, enzymology at 3, cell signaling at 6, and cytoskeleton and cancer are monotonically increasing with dehydron cluster size. These results also show that the cancer category becomes especially dominant for proteins that are very poorly wrapped. We identified domains having clusters of at least 7 dehydrons reported in the PDB. There were only 5 domains with n = 7: calmodulin (PDB 1CDM; cell signaling[349]), actin (PDB 1ATN; cytoskeleton[350]), p53 DNA-binding domain (PDB 2GEQ; cancer[347], Figures 5-3 and 5-7) BRCT, the terminal repeat domain of breast cancer gene BRAC1 (PDB 1JNX; cancer[351]) and the cellular prion protein (PDB 1QM0; not categorized[352]). The group of n > 7 only contains 3

members: severin (PDB 1SRV; cytoskeleton[353]) and two oncogenic transcription factors with DNA-stabilizing induced fit, jun/fos[354] (PDB 1FOS) and myc/max[355] (PDB 1A93). All eight protein domains having unusually large dehydron clusters belong to proteins involved in many interactions[356,187]. Even if functionally diverse, all these proteins have a common feature: as soluble proteins, they all possess a major weakness in the hydration shell. These dehydration hot spots play different yet interrelated roles: i) promoter of protein associations (calmodulin, actin, severin), ii) dielectric modulator enhancing intermolecular electrostatic interactions (cancer-related transcription factors), and iii) a structural weakness promoting water attack on backbone hydrogen bonds with concurrent refolding leading to aggregation (cellular prion protein).

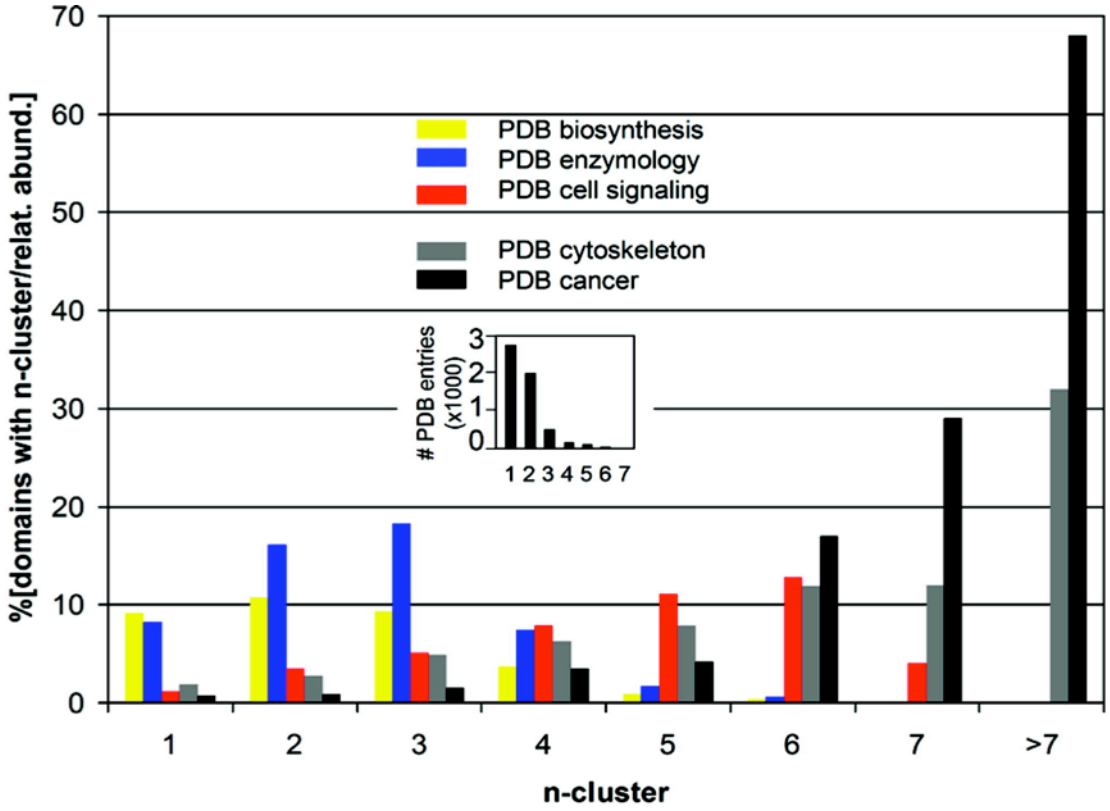

Figure 5-9. Percentages of PDB-domains in functional categories binned into groups determined by dehydron-cluster size n. Each cluster-size group is divided into five nondisjoint functional categories: biosynthesis, enzymology, cell signaling, cytoskeleton, and cancer. The number of PDB domains in each group is normalized to the relative abundance of the functional category. Thus, the number of PDB-domains in a cluster-size group and functional category is divided by the total number of PDB domains in the category. Inset: Number of domains in each cluster-size group.

5.5. Discussion

In this work, we inspected the defective packing of soluble proteins. We correlated insufficiently wrapped hydrogen bonds in protein backbones (i.e. dehydrons) with the level of confinement of hydrating water. By doing so, we

showed that while soluble proteins have a tight hydration shell and shielded intramolecular hydrogen bonds usually protecting them from water attack, proteins with deficiently packed hydrogen bonds exhibit a weakened hydration shell, thereby being capable of a local and discrete dielectric quenching. Specifically, we reported that for proteins with the largest clusters of dehydrons (7 or more) in the PDB, this capability of local and discrete quenching becomes more apparent. A search of the PDB revealed eight of such proteins having structural regions intermediate between order and disorder. These proteins included cancer-related proteins, highly interactive proteins and a cellular prion.

Our results suggest that protein regions in the twilight between order and disorder had several distinguishing features, such as being structurally unstable, having local dehydration propensity, and strong electrostatic enhancement on the protein surface, which may serve as functional indicators.

More recent work demonstrated that our approach - mainly founded on electrostatics and hydrogen bond complementarity - was appropriate for identifying protein regions with distinguishing properties making them especially suitable for interactions with other molecules. Chen and Lim, for example, identified the common physical basis for different macromolecular binding sites (i.e. DNA-,RNA- and protein binding) by considering fundamental principles of thermodynamics[50]. According to Chen and Lim, in the absence of their binding partner and solvent, residues involved in binding present suboptimal hydrogen-bonding interactions and packing, and are less energetically stable than non-macromolecule binding regions.

Similarly, Gonzalez-Ruiz and Gohlke reported on the role of water favoring the close packing of the atoms, which, in turn, ensures complementarity between binding partners. They also suggested that partial solvation was important for stabilizing charged groups in protein-protein interactions[357]. Currently, the importance of elucidating the fundamental biophysical principles that drive protein association is still topic of primary importance, since energetic determinants of affinity and specificity in protein interfaces are not fully understood. Disordered binding regions have lately been used as a proxy measure for protein interactions [184], while a very recent plug-in for the Molecular Graphics System PyMol (2012-01-14 Version 1.0), includes a dehydron calculator.

One of the major contributions of this work was that it established a connection between the concepts of dehydrons and intrinsic disorder. By exploring the PDB we related regions belonging to the order/disorder twilight in protein domains to their local dehydration propensity. We proposed that regions rich in packing defects were structurally unstable, making them candidate promoters of interactions.

Conclusions

This thesis approached the problem of identifying functional and structural features of protein sequences by implementing different strategies and focusing on features related to intrinsic disorder (ID).

We started with two studies that analyzed protein sequences by combining computational tools and expert knowledge, which proved their efficacy in guiding experimental assays. Potentially functionally relevant protein sites were computationally explored in a systematic manner, to be experimentally tested.

In the first study, we investigated the relationship between the sequence of the alpha synuclein (αsyn) protein and its aggregation propensity (AP). We identified three protein regions (hot spots) with predicted propensity to aggregate. Through systematic *in silico* mutagenesis of these hot spots based on reported physicochemical principles and previously reported experimental results, we designed αsyn protein variants with extreme predicted modulating effects on AP. Our results were consistent with previous observations: inserting charged residues[218] or aggregation breakers[211] decreased the predicted aggregation propensity (e.g. V71K and V71R; 54.4% decrease, V37P; 31.9% decrease). Introducing mutations increasing protein structure (i.e. reducing disorder content)[221,206,222] increased the predicted AP (e.g. H50V, H50I, 87% increase). We also confirmed the existence of several gatekeeper glycines which significantly

affect the protein aggregation propensity[209,219,220] (e.g. G14V, 52% inc.). In contrast to previous observations[226], our results showed that mutations disrupting known motifs of αsyn had little effect on the protein's aggregation propensity. Two of the proposed designed variants (V71K and V37P) experimentally demonstrated having solubility values in mammalian cells in agreement with our predictions (data not shown, LS and NP personal communications). Our strategy to identify AP hot spots, together with the rational design of protein variants modulating AP can be adopted to study other disease-associated aggregation-prone proteins. Understanding the mechanisms that govern aggregation can also have great impact in the biotechnological production and purification of recombinant proteins[243], which can also affect drug manufacturing[244]. Finally, the controlled self-assembly of proteins and peptides into aggregating structures may constitute an attractive alternative to develop nanomaterials[358].

In the second study, the goal was to provide a better characterization of the alanine racemase family according to the substrate specificity. The obtained family subdivision is far more accurate and consistent with recent observations on the substrate specificity of several members of the alanine racemase family[6] than current annotations. Using this classification, 77 alanine racemases from different bacterial species were identified as having a putatively broader specificity. Two of these newly classified enzymes have been crystalized, and their multi-binding specificity was experimentally confirmed (Data not shown, FC, personal communication). Interestingly, the resulting substrate specificity-determining

positions (SDPs) in the alanine racemase protein family included all the structurally and functionally relevant positions reported previously[238] and those provided by expert knowledge (FC, personal communication). We rationally designed point variants to experimentally test substrate binding, some of which were already experimentally tested (FC, personal communication, in preparation). The lack of alanine racemase function in eukaryotes[359] makes this enzyme an attractive target to develop novel antimicrobial drugs, especially because antibiotic resistance has become increasingly common over recent years[238,360]. Our work may also contribute to generate a better understanding of the production of noncanonical D-amino acids (NCDAAs). This knowledge can help further investigating how NCDAAs mediate distinct types of signals in mixed bacterial communities. Recently, NCDAAs have been shown to be used as a sort of "bacterial hormones", allowing the broadcasting of signals to the same bacteria/species and to other bacteria/species[6]. Additionally, NCDAAs have had an increasing application in the pharmaceutical industry, biotechnology, immunodiagnostics, and food industry in recent years[361,362,232].

In the second part of this thesis, we present a multi-scale assessment of the involvement of ID in protein function in different organisms, starting with the first genome-wide analysis of this phenomenon in *Arabidopsis thaliana*, then discussing its role in the specific process of vesicular trafficking in Human and yeast, and finally focusing on protein disorder at the atomic level of specific proteins of diverse organisms.

The analysis of *A. thaliana* revealed the functional classes enriched in intrinsically disordered proteins (IDPs). We found that proteins in functional classes related to environmental perception and response – fundamental for plant plasticity – showed enrichment in intrinsically disordered regions (IDRs) with respect to Human. These results are consistent with previous observations on the relationship between ID and processes involved in response to environmental stimuli[246,124,125,126]. Furthermore, our findings fit the notion that newly introduced IDPs and IDRs serve mainly as carries for new binding regions in eukaryotic organisms[85], thus adding complexity to the system. ID increases the complexity of biological processes and protein networks by increasing their "wiring" (e.g. the potential connections between proteins), and this increased complexity is especially evident in those protein networks underlying phenotypic plasticity and adaptation to environmental stress. In addition, IDPs/IDRs' tolerance for mutations allows them to undergo fast evolutionary changes[115,116,117], thus providing exceptional adaptability. Thus, our results support the correlation between complexity and protein disorder, and suggest that plants have used ID as an evolutionary tool to increase complexity and adaptability in their biological networks.

The systematic assessment of ID in the main vesicle trafficking pathways in Human and yeast confirmed that proteins in the clathrin system are significantly more disordered than proteins in the COPI and COPII systems. This enrichment in ID may partly account for the presence of highly dynamic protein networks in the clathirn route, which is agreement with the highest versatility of this route[324,325].

Similarly, ID is likely responsible for the observed high evolutionary plasticity and robustness of clathrin-associated proteins, exhibiting many species-specific characteristics[288,292], yet extensively modified to assist other specialized pathways in many organisms[328,329,330]. This work confirmed that ID is widespread and frequently essential for proteins involved in vesicle trafficking. Additionally, ID differential abundance patterns among the different routes provided the structural background for long standing observations on the functional and evolutionary differences of these vesicle trafficking systems.

In the last part of this work, we show the seminal work of this dissertation. This work connected the concept of dehydrons[363] with intrinsic disorder. We related regions belonging to the order/disorder twilight in protein domains to their local dehydration propensity. We confirmed that regions rich in packing defects were structurally unstable and promoters of interactions. One of the main contributions of this work was to advance two fundamental and related notions in the ID field: the observation that protein regions may become folded (stabilized) upon binding, and that as a consequence, they foster molecular interactions. We also proposed that unstable protein regions may be related to diseases due to their presence in highly connected proteins. These two observations stood the test of time, since it is now beyond doubt that IDPs/IDRs act as hubs in many key pathways[16,133,142] and as such are commonly implicated in many diseases too[14,12,8].

As a final remark, we want to address some fundamental issues regarding the study of ID. From the computational perspective, ID assessment should be

performed in a more systematic way than what current approaches implement. Adopting the same disorder prediction measurements (e.g. ratio of disordered residues to ordered residues, number of long disordered windows, number of disordered binding sites) will help providing a more coherent picture of ID's abundance and functions among different pathways and organisms. This standardization in ID measurement should also consider the two major functional distinctions of IDRs: i) linking or spacing between domains, and ii) molecular recognition. Thus, different ID measurements should be adopted for capturing specific ID roles. Moreover, if this functional distinction were adopted when implementing disorder predictors, their accuracy would likely improve. Making this refinement in the methods probably result in a better agreement among their prediction, as well as between predictions and experimental data.

From an experimental point of view, there is an impeding need to understand IDP regulation in the context of living cells. *In vivo* experiments are crucial in understanding how IDPs exist, persist and function in cells. However, this requires the development of structural techniques capable of determining the ensemble of structures of proteins and capturing them as they twist and turn in solution.

The collection of structural–functional studies of the many IDPs and IDRs presented in this thesis contribute to a better understanding of this phenomenon in different organisms and biological processes. Additionally, these results provide evidence of the use of ID as a mechanism to increase the complexity of protein and biological networks, and as a means to increase the adaptability of proteins in

specific processes. Thus, our results contribute to elucidate the relationship between network and organismal complexity and ID, while they also provide evidence of the evolutionary advantages offered by ID.

References

1. Pietrosemoli, N., Lopez, D., Segura-Cabrera, A. & Pazos, F. Computational Prediction of Important Regions in Protein Sequences [Life Sciences]. *IEEE Signal Processing Magazine* **29**, 143 – 147 (2012).

2. Uversky, V. N. & Eliezer, D. Biophysics of Parkinson's disease: structure and aggregation of alpha-synuclein. *Curr. Protein Pept. Sci.* **10**, 483–499 (2009).

3. de Lau, L. M. L. & Breteler, M. M. B. Epidemiology of Parkinson's disease. *Lancet Neurol* **5**, 525–535 (2006).

4. Cookson, M. R. alpha-Synuclein and neuronal cell death. *Mol Neurodegener* **4**, 9 (2009).

5. Tanner, M. E. Understanding nature's strategies for enzyme-catalyzed racemization and epimerization. *Acc. Chem. Res.* **35**, 237–246 (2002).

6. Lam, H. *et al.* D-amino acids govern stationary phase cell wall remodeling in bacteria. *Science* **325**, 1552–1555 (2009).

7. Gsponer, J. & Madan Babu, M. The rules of disorder or why disorder rules. *Progress in Biophysics and Molecular Biology* **99**, 94–103 (2009).

8. Dunker, A. K. *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9**, S1 (2008).

9. Dyson, H. J. Expanding the proteome: disordered and alternatively folded proteins. *Q. Rev. Biophys.* **44**, 467–518 (2011).

10. Xie, H. *et al.* Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* **6**, 1882 – 1898 (2007).

11. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).

12. Uversky, V. N. Amyloidogenesis of natively unfolded proteins. *Current Alzheimer Research* **5**, 260–287 (2008).

13. Breydo, L., Wu, J. W. & Uversky, V. N. α-Synuclein misfolding and Parkinson's disease. *Biochim. Biophys. Acta* **1822**, 261–285 (2012).

14. Uversky, V. N. Neuropathology, biochemistry, and biophysics of alpha-synuclein aggregation. *J Neurochem* **103**, 17 – 37 (2007).

15. Uversky, V. N. A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *J Biomol Struct Dyn* **21**, 211 – 234 (2003).

16. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., ObradoviÄ‡, Z. & Dunker, A. K. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *Journal of Molecular Biology* **323**, 573–584 (2002).

17. Metallo, S. J. Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol* **14**, 481–488 (2010).

18. Tompa, P. Unstructural biology coming of age. *Curr. Opin. Struct. Biol* **21**, 419–425 (2011).

19. Uversky, V. N. & Dunker, A. K. Multiparametric Analysis of Intrinsically Disordered Proteins: Looking at Intrinsic Disorder through Compound Eyes. *Anal. Chem.* **84**, 2096–2104 (2011).

20. Casal, J. J., Fankhauser, C., Coupland, G. & Blázquez, M. A. Signalling for developmental plasticity. *Trends Plant Sci.* **9**, 309–314 (2004).

21. Pietrosemoli, N., Crespo, A. & Fernandez, A. Dehydration propensity of order-disorder intermediate regions in soluble proteins. *J. Proteome Res.* **6**, 3519–3526 (2007).

22. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496–512 (1995).

23. Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40**, D571–579 (2012).

24. The UniProt Consortium Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **40**, D71–D75 (2011).

25. Panchenko, A. R., Kondrashov, F. & Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci* **13**, 884–892 (2004).

26. Pazos, F. & Bang, J.-W. Computational Prediction of Functionally Important Regions in Proteins. *Current Bioinformatics* **1**, 15–23 (2006).

27. Berman, H. M. *et al.* The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242 (2000).

28. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **8**, 995–1005 (2007).

29. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113 (1970).

30. Frasconi, P. & Shamir, R. *Artificial Intelligence and Heuristic Methods in Bioinformatics*. (IOS Press, 2003).

31. Zuckerkandl, E. & Pauling, L.in *Evolving Genes and Proteins* 97–166 (Academic Press, 1965).

32. Ouzounis, C., Pérez-Irratxeta, C., Sander, C. & Valencia, A. Are binding residues conserved? *Pac Symp Biocomput* 401–412 (1998).

33. Valdar, W. S. & Thornton, J. M. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42**, 108–124 (2001).

34. Pei, J. & Grishin, N. V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**, 700–712 (2001).

35. Armon, A., Graur, D. & Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology* **307**, 447–463 (2001).

36. Mirny, L. A. & Gelfand, M. S. Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors. *Journal of Molecular Biology* **321**, 7–20 (2002).

37. Kalinina, O. V., Novichkov, P. S., Mironov, A. A., Gelfand, M. S. & Rakhmaninova, A. B. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.* **32**, W424–428 (2004).

38. Donald, J. E. & Shakhnovich, E. I. Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucleic Acids Research* **33**, 4455 –4465

39. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).

40. Andrade, M. A., Casari, G., Sander, C. & Valencia, A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol Cybern* **76**, 441–450 (1997).

41. Bauer, B. *et al.* Effector Recognition by the Small GTP-binding Proteins Ras and Ral. *Journal of Biological Chemistry* **274**, 17763 –17770 (1999).

42. Tress, M. *et al.* Scoring docking models with evolutionary information. *Proteins: Structure, Function, and Bioinformatics* **60**, 275–280 (2005).

43. Mihalek, I., Res, I. & Lichtarge, O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.* **336**, 1265–1282 (2004).

44. Rausell, A., Juan, D., Pazos, F. & Valencia, A. Protein interactions and ligand binding: From protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A* **107**, 1995–2000 (2010).

45. Muth, T. *et al.* JDet: interactive calculation and visualization of function-related conservation patterns in multiple sequence alignments and structures. *Bioinformatics* **28**, 584–586 (2012).

46. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl 1**, S71–77 (2002).

47. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408 (2001).

48. Landgraf, R., Xenarios, I. & Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502 (2001).

49. Pettit, F. K., Bare, E., Tsai, A. & Bowie, J. U. HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.* **369**, 863–879 (2007).

50. Chen, Y. C. & Lim, C. Common physical basis of macromolecule-binding sites in proteins. *Nucleic Acids Res* **36**, 7078–7087 (2008).

51. Grosdidier, S. & Fernández-Recio, J. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics* **9**, 447 (2008).

52. Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins* **66**, 630–645 (2007).

53. Dunker, A. K. *et al.* Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput* 473 – 484 (1998).

54. Garner, Cannon, Romero, Obradovic & Dunker Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization. *Genome Inform Ser Workshop Genome Inform* **9**, 201–213 (1998).

55. Romero, P. *et al.* Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 437 – 448 (1998).

56. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* **293**, 321–331 (1999).

57. Sedzik, J. & Kirschner, D. A. Is myelin basic protein crystallizable? *Neurochem. Res.* **17**, 157–166 (1992).

58. Pauling, L. A Theory of the Structure and Process of Formation of Antibodies*. *J. Am. Chem. Soc.* **62**, 2643–2657 (1940).

59. Williams, R. J. The conformational mobility of proteins and its functional significance. *Biochem. Soc. Trans.* **6**, 1123–1126 (1978).

60. Romero, P. *et al.* Sequence complexity of disordered protein. *Proteins* **42**, 38 – 48 (2001).

61. Sickmeier, M. *et al.* DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* D786 – 793 (2007).

62. Melamud, E. & Moult, J. Evaluation of disorder predictions in CASP5. *Proteins: Structure, Function, and Bioinformatics* **53**, 561–565 (2003).

63. Jin, Y. & Dunbrack, R. L. Assessment of disorder predictions in CASP6. *Proteins: Structure, Function, and Bioinformatics* **61**, 167–175 (2005).

64. Bordoli, L., Kiefer, F. & Schwede, T. Assessment of disorder predictions in CASP7. *Proteins: Structure, Function, and Bioinformatics* **69**, 129–136 (2007).

65. Noivirt-Brik, O., Prilusky, J. & Sussman, J. L. Assessment of disorder predictions in CASP8. *Proteins* **77**, 210–216 (2009).

66. Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A. & Kryshtafovych, A. Evaluation of disorder predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics* **79**, 107–118 (2011).

67. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1804**, 1231–1264 (2010).

68. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11**, 161 – 171 (2000).

69. Warnefors, M. & Eyre-Walker, A. The Accumulation of Gene Regulation Through Time. *Genome Biology and Evolution* **3**, 667 –673 (2011).

70. Dunker, A. K. & Obradovic, Z. The protein trinity-linking function and disorder. *Nat Biotech* **19**, 805–806 (2001).

71. Dosztanyi, Z., Chen, J., Dunker, A. K., Simon, I. & Tompa, P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* **5**, 2985 – 2995 (2006).

72. Haynes, C. *et al.* Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* **2**, e100 (2006).

73. Hsu, W.-L. *et al.* Intrinsic protein disorder and protein-protein interactions. *Pac Symp Biocomput* 116–127 (2012).

74. Tompa, P., Szász, C. & Buday, L. Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.* **30**, 484–489 (2005).

75. Tompa, P., Dosztanyi, Z. & Simon, I. Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J. Proteome Res.* **5**, 1996–2000 (2006).

76. Bellay, J. *et al.* Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* **12**, R14 (2011).

77. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197 – 208 (2005).

78. Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S. & Dunker, A. K.in *Protein Folding Handbook* (Buchner, J. & Kiefhaber, T.) 275–357 (Wiley-VCH Verlag GmbH, 2008).at <http://onlinelibrary.wiley.com/doi/10.1002/9783527619498.ch41/summary>

79. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**, 739 – 756 (2002).

80. Galzitskaya, O. V., Bogatyreva, N. S. & Ivankov, D. N. Compactness determines protein folding type. *J Bioinform Comput Biol* **6**, 667–680 (2008).

81. Turoverov, K. K., Kuznetsova, I. M. & Uversky, V. N. The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation. *Prog. Biophys. Mol. Biol.* **102**, 73–84 (2010).

82. Peter, T. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters* **579**, 3346–3354 (2005).

83. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573 – 6582 (2002).

84. Nagy, A. *et al.* Hierarchical Extensibility in the PEVK Domain of Skeletal-Muscle Titin. *Biophys J* **89**, 329–336 (2005).

85. Dosztányi, Z., Sándor, M., Tompa, P. & Simon, I. Prediction of protein disorder at the domain level. *Curr. Protein Pept. Sci* **8**, 161–171 (2007).

86. Mukhopadhyay, R. & Hoh, J. H. AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. *FEBS Lett.* **505**, 374–378 (2001).

87. Melissa M Pentony, Ward, J. & Jones, D. T.in *Proteome Bioinformatics* **604**, 369–393 (Springer, 2010).

88. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527 – 533 (2002).

89. Tompa, P. & Csermely, P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* **18**, 1169–1175 (2004).

90. Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037–1049 (2004).

91. Cox, C. J. *et al.* The regions of securin and cyclin B proteins recognized by the ubiquitination machinery are natively unfolded. *FEBS Lett.* **527**, 303–308 (2002).

92. Davies, K. J. Degradation of oxidized proteins by the 20S proteasome. *Biochimie* **83**, 301–310 (2001).

93. David, D. C. *et al.* Proteasomal degradation of tau protein. *Journal of Neurochemistry* **83**, 176–185 (2002).

94. Sheaff, R. J. *et al.* Proteasomal turnover of p21Cip1 does not require p21Cip1 ubiquitination. *Mol. Cell* **5**, 403–410 (2000).

95. Olashaw, N., Bagui, T. K. & Pledger, W. J. Cell cycle control: a complex issue. *Cell Cycle* **3**, 263–264 (2004).

96. Shoemaker, B. A., Portman, J. J. & Wolynes, P. G. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 8868–8873 (2000).

97. Spolar, R. S. & Record, M. T., Jr Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777–784 (1994).

98. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026 (2004).

99. Wright, P. E. & Dyson, H. J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **19**, 31–38 (2009).

100. Dunker, A. K. *et al.* Intrinsically disordered protein. *J Mol Graph Model* **19**, 26 – 59 (2001).

101. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I. & Wright, P. E. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 11504–11509 (1996).

102. Oldfield, C. J. *et al.* Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9**, S1 (2008).

103. Mohan, A. *et al.* Analysis of molecular recognition features (MoRFs). *J Mol Biol* **362**, 1043 – 1059 (2006).

104. Oldfield, C. J. *et al.* Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**, 12454 – 12470 (2005).

105. Cheng, Y., Oldfield, C. J., Romero, P., Uversky, V. N. & Dunker, A. K. Mining alpha-helix-forming molecular recognition features alpha-MoRFs with cross species sequence alignments. *Biochemistry* **46**, 13468 – 13477 (2007).

106. Gould, C. M. *et al.* ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res* **38**, D167–D180 (2010).

107. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746 (2009).

108. Tompa, P. *et al.* Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* **31**, 328–335 (2009).

109. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–141 (2004).

110. Neduva, V. & Russell, R. B. Linear motifs: evolutionary interaction switches. *FEBS Lett.* **579**, 3342–3345 (2005).

111. Davey, N. E., Shields, D. C. & Edwards, R. J. SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res.* **34**, 3546–3554 (2006).

112. Ponting, C. P., Schultz, J., Copley, R. R., Andrade, M. A. & Bork, P. Evolution of domain families. *Adv. Protein Chem.* **54**, 185–244 (2000).

113. Dayhoff, M. O. & Schwartz, R. M. Chapter 22: A model of evolutionary change in proteins. in *in Atlas of Protein Sequence and Structure* (1978).

114. Brown, C. J., Johnson, A. K., Dunker, A. K. & Daughdrill, G. W. Evolution and disorder. *Current Opinion in Structural Biology* **21**, 441–446 (2011).

115. Daughdrill, G. W., Narayanaswami, P., Gilmore, S. H., Belczyk, A. & Brown, C. J. Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. *J. Mol. Evol.* **65**, 277–288 (2007).

116. Denning, D. P. & Rexach, M. F. Rapid evolution exposes the boundaries of domain structure and function in natively unfolded FG nucleoporins. *Mol. Cell Proteomics* **6**, 272–282 (2007).

117. Ayme-Southgate, A. J., Southgate, R. J., Philipp, R. A., Sotka, E. E. & Kramp, C. The myofibrillar protein, projectin, is highly conserved across insect evolution except for its PEVK domain. *J. Mol. Evol.* **67**, 653–669 (2008).

118. Chen, J. W., Romero, P., Uversky, V. N. & Dunker, A. K. Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder. *J Proteome Res* **5**, 888–898 (2006).

119. Chen, J. W., Romero, P., Uversky, V. N. & Dunker, A. K. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J. Proteome Res* **5**, 879–887 (2006).

120. Brown, C. J., Johnson, A. K. & Daughdrill, G. W. Comparing Models of Evolution for Ordered and Disordered Proteins. *Mol Biol Evol* **27**, 609–621 (2010).

121. Szalkowski, A. M. & Anisimova, M. Markov Models of Amino Acid Substitution to Study Proteins with Intrinsically Disordered Regions. *PLoS ONE* **6**, e20488 (2011).

122. Chen, S. C.-C., Chen, F.-C. & Li, W.-H. Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals. *Mol. Biol. Evol.* **27**, 2548–2554 (2010).

123. Mosca, R., Pache, R. A. & Aloy, P. The role of structural disorder in the rewiring of protein interactions through evolution. *Molecular & Cellular Proteomics: MCP* (2012).doi:10.1074/mcp.M111.014969

124. Schlessinger, A. *et al.* Protein disorder-a breakthrough invention of evolution? *Curr. Opin. Struct. Biol* **21**, 412–418 (2011).

125. Xue, B., Williams, R. W., Oldfield, C. J., Dunker, A. K. & Uversky, V. N. Archaic chaos: intrinsically disordered proteins in Archaea. *BMC Syst Biol* **4 Suppl 1**, S1 (2010).

126. Awile, O., Krisko, A., Sbalzarini, I. F. & Zagrovic, B. Intrinsically Disordered Regions May Lower the Hydration Free Energy in Proteins: A Case Study of Nudix Hydrolase in the Bacterium Deinococcus radiodurans. *PLoS Comput Biol* **6**, e1000854 (2010).

127. Andreeva, A. & Murzin, A. G. Structural classification of proteins and structural genomics: new insights into protein folding and evolution. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **66**, 1190–1197 (2010).

128. Murzin, A. G. Metamorphic Proteins. *Science* **320**, 1725–1726 (2008).

129. Uversky, V. N. Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. *Chem. Soc. Rev.* **40**, (2010).

130. He, B. *et al.* Predicting intrinsic disorder in proteins: an overview. *Cell Res* **19**, 929–949 (2009).

131. Sugase, K., Dyson, H. J. & Wright, P. E. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **447**, 1021–1025 (2007).

132. Liu, J., Faeder, J. R. & Camacho, C. J. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proceedings of the National Academy of Sciences* **106**, 19819 – 19823 (2009).

133. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs J* **272**, 5129 – 5148 (2005).

134. Patil, A. & Nakamura, H. Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett* **580**, 2041 – 2045 (2006).

135. Kim, P. M., Sboner, A., Xia, Y. & Gerstein, M. The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol* **4**, (2008).

136. Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).

137. Li, S. *et al.* A Map of the Interactome Network of the Metazoan C. elegans. *Science* **303**, 540–543 (2004).

138. Giot, L. *et al.* A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727–1736 (2003).

139. Barrios-Rodiles, M. *et al.* High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307**, 1621–1625 (2005).

140. Kim, P. M., Sboner, A., Xia, Y. & Gerstein, M. The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol* **4**, (2008).

141. Schlessinger, A., Liu, J. & Rost, B. Natively Unstructured Loops Differ from Other Loops. *PLoS Comput Biol* **3**, e140 (2007).

142. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* **18**, 343 – 384 (2005).

143. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci* **33**, 2–8 (2008).

144. Fontes, M. R. M. *et al.* Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-alpha. *Biochem J* **375**, 339–349 (2003).

145. von Ossowski, I. *et al.* Protein Disorder: Conformational Distribution of the Flexible Linker in a Chimeric Double Cellulase. *Biophys J* **88**, 2823–2832 (2005).

146. Rock, R. S. *et al.* A flexible domain is essential for the large step size and processivity of myosin VI. *Mol. Cell* **17**, 603–609 (2005).

147. Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA processing with transcription. *Cell* **108**, 501–512 (2002).

148. Hazy, E. & Tompa, P. Limitations of induced folding in molecular recognition by intrinsically disordered proteins. *Chemphyschem* **10**, 1415–1419 (2009).

149. Mittag, T. *et al.* Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc Natl Acad Sci U S A* **105**, 17772–17777 (2008).

150. Mittag, T., Kay, L. E. & Forman-Kay, J. D. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* **23**, 105–116 (2010).

151. Permyakov, S. E., Millett, I. S., Doniach, S., Permyakov, E. A. & Uversky, V. N. Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. *Proteins* **53**, 855–862 (2003).

152. Mittag, T. *et al.* Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **18**, 494–506 (2010).

153. Sigalov, A., Aivazian, D. & Stern, L. Homooligomerization of the cytoplasmic domain of the T cell receptor zeta chain and of other proteins containing the immunoreceptor tyrosine-based activation motif. *Biochemistry* **43**, 2049 – 2061 (2004).

154. Sigalov, A. B., Zhuravleva, A. V. & Orekhov, V. Y. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. *Biochimie* **89**, 419 – 421 (2007).

155. Pometun, M. S., Chekmenev, E. Y. & Wittebort, R. J. Quantitative observation of backbone disorder in native elastin. *J. Biol. Chem.* **279**, 7982–7987 (2004).

156. Sigalov, A. B. & Hendricks, G. M. Membrane binding mode of intrinsically disordered cytoplasmic domains of T cell receptor signaling subunits depends on lipid composition. *Biochem. Biophys. Res. Commun.* **389**, 388–393 (2009).

157. Sigalov, A. B., Aivazian, D. A., Uversky, V. N. & Stern, L. J. Lipid-binding activity of intrinsically unstructured cytoplasmic domains of multichain immune recognition receptor signaling subunits. *Biochemistry* **45**, 15731 – 15739 (2006).

158. Salmon, L. *et al.* NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **132**, 8407–8418 (2010).

159. Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. Protein disk of tobacco mosaic virus at 2.8 A resolution showing the interactions within and between subunits. *Nature* **276**, 362–368 (1978).

160. Bode, W., Schwager, P. & Huber, R. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 A resolution. *J. Mol. Biol.* **118**, 99–112 (1978).

161. Binolfi, A., Theillet, F.-X. & Selenko, P. Bacterial in-cell NMR of human α-synuclein: a disordered monomer by nature? *Biochem. Soc. Trans.* **40**, 950–954 (2012).

162. Ito, Y., Mikawa, T. & Smith, B. O. In-cell NMR of intrinsically disordered proteins in prokaryotic cells. *Methods Mol. Biol.* **895**, 19–31 (2012).

163. Wang, G.-F., Li, C. & Pielak, G. J. 19F NMR studies of α-synuclein-membrane interactions. *Protein Sci.* **19**, 1686–1691 (2010).

164. Bodart, J.-F. *et al.* NMR observation of Tau in *Xenopus* oocytes. *Journal of Magnetic Resonance* **192**, 252–257 (2008).

165. Kodera, N., Yamamoto, D., Ishikawa, R. & Ando, T. Video imaging of walking myosin V by high-speed atomic force microscopy. *Nature* **468**, 72–76 (2010).

166. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E. & Dunker, A. K. Identifying disordered regions in proteins from amino acid sequence. *1997 Proceedings of International Conference on Neural Networks* **1**, 90 – 95 (1997).

167. Dosztányi, Z., Mészáros, B. & Simon, I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Briefings in Bioinformatics* **11**, 225 –243 (2010).

168. Prilusky, J. *et al.* FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435–3438 (2005).

169. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research* **31**, 3701 –3708 (2003).

170. Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).

171. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635 – 645 (2004).

172. Vullo, A., Bortolami, O., Pollastri, G. & Tosatto, S. C. E. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.* **34**, W164–168 (2006).

173. in *Structure and Function of Intrinsically Disordered Proteins* (Chapman and Hall/CRC, 2009).at <http://www.crcnetbase.com/doi/abs/10.1201/9781420078930.fmatt>

174. Dosztányi, Z., Fiser, A. & Simon, I. Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.* **272**, 597–612 (1997).

175. Galzitskaya, O. V., Garbuzynskiy, S. O. & Lobanov, M. Y. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* **22**, 2948–2949 (2006).

176. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).

177. Bujnicki, J. M., Elofsson, A., Fischer, D. & Rychlewski, L. LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins* **Suppl 5**, 184–191 (2001).

178. Orosz, F. & Ovádi, J. Proteins without 3D structure: definition, detection and beyond. *Bioinformatics* **27**, 1449 –1454 (2011).

179. Ishida, T. & Kinoshita, K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* **24**, 1344–1348 (2008).

180. PLoS ONE: Improved Disorder Prediction by Combination of Orthogonal Approaches. at <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004433>

181. Lieutaud, P., Canard, B. & Longhi, S. MeDor: a metaserver for predicting protein disorder. *BMC Genomics* **9**, S25 (2008).

182. Hegyi, H., Schad, E. & Tompa, P. Structural disorder promotes assembly of protein complexes. *BMC Struct Biol* **7**, 65–65

183. Li, Romero, Rani, Dunker & Obradovic Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform* **10**, 30–40 (1999).

184. Buljan, M. *et al.* Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Mol. Cell* **46**, 871–883 (2012).

185. Gsponer, J., Futschik, M. E., Teichmann, S. A. & Babu, M. M. Tight regulation of unstructured proteins. *Science* **322**, 1365–1368 (2008).

186. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology* **21**, 432–440 (2011).

187. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

188. Uversky, V. N. What does it mean to be natively unfolded? *Eur. J. Biochem.* **269**, 2–12 (2002).

189. Vavouri, T., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198–208 (2009).

190. Cheng, Y. *et al.* Rational drug design via intrinsically disordered protein. *Trends Biotechnol.* **24**, 435–442 (2006).

191. Vassilev, L. T. *et al.* In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848 (2004).

192. Hammoudeh, D. I., Follis, A. V., Prochownik, E. V. & Metallo, S. J. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. *J. Am. Chem. Soc.* **131**, 7390–7401 (2009).

193. Srinivasan, M. & Dunker, A. K. Proline rich motifs as drug targets in immune mediated disorders. *Int J Pept* **2012**, 634769 (2012).

194. Castillo, V. & Ventura, S. Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. *PLoS Comput. Biol.* **5**, e1000476 (2009).

195. Kothawala, A., Kilpatrick, K., Novoa, J. A. & Segatori, L. Quantitative analysis of α-synuclein solubility in living cells using split GFP complementation. *PLoS ONE* **7**, e43505 (2012).

196. Karpinar, D. P. *et al.* Pre-fibrillar alpha-synuclein variants with impaired beta-structure increase neurotoxicity in Parkinson's disease models. *EMBO J.* **28**, 3256–3268 (2009).

197. Lashuel, H. A. & Lansbury, P. T., Jr Are amyloid diseases caused by protein aggregates that mimic bacterial pore-forming toxins? *Q. Rev. Biophys.* **39**, 167–201 (2006).

198. Ventura, S. Sequence determinants of protein aggregation: tools to increase protein solubility. *Microb. Cell Fact.* **4**, 11 (2005).

199. Ventura, S. *et al.* Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7258–7263 (2004).

200. Castillo, V., Graña-Montes, R., Sabate, R. & Ventura, S. Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. *Biotechnol J* **6**, 674–685 (2011).

201. Pawlicki, S., Le Béchec, A. & Delamarche, C. AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics* **9**, 273 (2008).

202. Kumar, S., Sarkar, A. & Sundar, D. Controlling aggregation propensity in A53T mutant of alpha-synuclein causing Parkinson's disease. *Biochem. Biophys. Res. Commun.* **387**, 305–309 (2009).

203. Conway, K. A. *et al.* Acceleration of oligomerization, not fibrillization, is a shared property of both α-synuclein mutations linked to early-onset Parkinson's disease: Implications for pathogenesis and therapy. *PNAS* **97**, 571–576 (2000).

204. Greenbaum, E. A. *et al.* The E46K mutation in alpha-synuclein increases amyloid fibril formation. *J. Biol. Chem.* **280**, 7800–7807 (2005).

205. Zarranz, J. J. *et al.* The new mutation, E46K, of alpha-synuclein causes Parkinson and Lewy body dementia. *Ann Neurol* **55**, 164 – 173 (2004).

206. Beyer, K. Alpha-synuclein structure, posttranslational modification and alternative splicing as aggregation enhancers. *Acta Neuropathol.* **112**, 237–251 (2006).

207. Beyer, K. *et al.* Identification and characterization of a new alpha-synuclein isoform and its role in Lewy body diseases. *Neurogenetics* **9**, 15–23 (2008).

208. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).

209. Rousseau, F., Schymkowitz, J. & Serrano, L. Protein aggregation and amyloidosis: confusion of the kinds? *Curr. Opin. Struct. Biol.* **16**, 118–126 (2006).

210. Maurer-Stroh, S. *et al.* Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods* **7**, 237–242 (2010).

211. Conchillo-Solé, O. *et al.* AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics* **8**, 65 (2007).

212. Tartaglia, G. G. *et al.* Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.* **380**, 425–436 (2008).

213. Pawar, A. P. *et al.* Prediction of 'aggregation-prone' and 'aggregation-susceptible' regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.* **350**, 379–392 (2005).

214. DuBay, K. F. *et al.* Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* **341**, 1317–1326 (2004).

215. Tartaglia, G. G. & Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev* **37**, 1395–1401 (2008).

216. Linding, R., Schymkowitz, J., Rousseau, F., Diella, F. & Serrano, L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342**, 345–353 (2004).

217. Linding, R., Schymkowitz, J., Rousseau, F., Diella, F. & Serrano, L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342**, 345–353 (2004).

218. Calloni, G., Zoffoli, S., Stefani, M., Dobson, C. M. & Chiti, F. Investigating the effects of mutations on protein aggregation in the cell. *J. Biol. Chem.* **280**, 10607–10613 (2005).

219. Parrini, C. *et al.* Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure* **13**, 1143–1151 (2005).

220. Fowler, S. B. *et al.* Rational design of aggregation-resistant bioactive peptides: Reengineering human calcitonin. *PNAS* **102**, 10105–10110 (2005).

221. Uversky, V. N., Li, J. & Fink, A. L. Evidence for a partially folded intermediate in alpha-synuclein fibril formation. *J. Biol. Chem.* **276**, 10737–10744 (2001).

222. McLean, P. J. & Hyman, B. T. An alternatively spliced form of rodent alpha-synuclein forms intracellular inclusions in vitro: role of the carboxy-terminus in alpha-synuclein aggregation. *Neurosci. Lett.* **323**, 219–223 (2002).

223. Uversky, V. N. Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front. Biosci.* **14**, 5188–5238 (2009).

224. Beyer, K. α-Synuclein structure, posttranslational modification and alternative splicing as aggregation enhancers. *Acta Neuropathologica* **112**, 237–251 (2006).

225. Waxman, E. A., Mazzulli, J. R. & Giasson, B. I. Characterization of hydrophobic residue requirements for alpha-synuclein fibrillization. *Biochemistry* **48**, 9427–9436 (2009).

226. Sode, K. Effect of Reparation of Repeat Sequences in the Human α-Synuclein on Fibrillation Ability. *International Journal of Biological Sciences* 1 (2007).doi:10.7150/ijbs.3.1

227. Chang, Y. F. & Adams, E. D-lysine catabolic pathway in Pseudomonas putida: interrelations with L-lysine catabolism. *J. Bacteriol.* **117**, 753–764 (1974).

228. Conrad, R. S., Massey, L. K. & Sokatch, J. R. D- and L-isoleucine metabolism and regulation of their pathways in Pseudomonas putida. *J. Bacteriol.* **118**, 103–111 (1974).

229. Pioli, D., Venables, W. A. & Franklin, F. C. D-Alanine dehydrogenase. Its role in the utilisation of alanine isomers as growth substrates by Pseudomonas aeruginosa PA01. *Arch. Microbiol.* **110**, 287–293 (1976).

230. Horcajo, P., de Pedro, M. A. & Cava, F. Peptidoglycan Plasticity in Bacteria: Stress-Induced Peptidoglycan Editing by Noncanonical D-Amino Acids. *Microbial Drug Resistance* **18**, 306–313 (2012).

231. Dimmer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic acids research* **40**, D565–570 (2012).

232. Friedman, M. Origin, Microbiology, Nutrition, and Pharmacology of D-Amino Acids. *Chemistry & Biodiversity* **7**, 1491–1530 (2010).

233. Pazos, F., Rausell, A. & Valencia, A. Phylogeny-independent detection of functional residues. *Bioinformatics* **22**, 1440–1448 (2006).

234. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389 – 3402 (1997).

235. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792–1797 (2004).

236. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

237. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research* **39**, W475–W478 (2011).

238. Im, H., Sharpe, M., Strych, U., Davlieva, M. & Krause, K. The crystal structure of alanine racemase from Streptococcus pneumoniae, a target for structure-based drug design. *BMC Microbiology* **11**, 116 (2011).

239. Stamper, G. F., Morollo, A. A., Ringe, D. & Stamper, C. G. Reaction of alanine racemase with 1-aminoethylphosphonic acid forms a stable external aldimine. *Biochemistry* **37**, 10438–10445 (1998).

240. Ondrechen, M. J., Briggs, J. M. & McCammon, J. A. A model for enzyme-substrate interaction in alanine racemase. *J. Am. Chem. Soc.* **123**, 2830–2834 (2001).

241. Watanabe, A., Yoshimura, T., Mikami, B. & Esaki, N. Tyrosine 265 of alanine racemase serves as a base abstracting alpha-hydrogen from L-alanine: the counterpart residue to lysine 39 specific to D-alanine. *J. Biochem.* **126**, 781–786 (1999).

242. LeMagueres, P. *et al.* The 1.9 A crystal structure of alanine racemase from Mycobacterium tuberculosis contains a conserved entryway into the active site. *Biochemistry* **44**, 1471–1481 (2005).

243. Ventura, S. & Villaverde, A. Protein quality in bacterial inclusion bodies. *Trends Biotechnol.* **24**, 179–185 (2006).

244. Rosenberg, A. S. Effects of protein aggregates: an immunologic perspective. *AAPS J* **8**, E501–507 (2006).

245. Morillas, M. *et al.* Identification of Conserved Amino Acid Residues in Rat Liver Carnitine Palmitoyltransferase I Critical for Malonyl-CoA Inhibition MUTATION OF METHIONINE 593 ABOLISHES MALONYL-CoA INHIBITION. *J. Biol. Chem.* **278**, 9058–9063 (2003).

246. Schad, E., Tompa, P. & Hegyi, H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biology* **12**, R120 (2011).

247. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins in human diseases: Introducing the D2 concept. *Ann Rev Biophys Biomol Structure* **37**, 215 – 246 (2008).

248. Pentony, M. M. & Jones, D. T. Modularity of intrinsic disorder in the human proteome. *Proteins* **78**, 212–221 (2010).

249. Yruela, I. & Contreras-Moreira, B. Protein disorder in plants: a view from the chloroplast. *BMC Plant Biol.* **12**, 165 (2012).

250. Sun, X. *et al.* A functionally required unfoldome from the plant kingdom: intrinsically disordered N-terminal domains of GRAS proteins are involved in molecular recognition during plant development. *Plant Mol. Biol.* **77**, 205–223 (2011).

251. Tompa, P. & Kovacs, D. Intrinsically disordered chaperones in plants and animals. *Biochem. Cell Biol.* **88**, 167–174 (2010).

252. Mouillon, J.-M., Gustafsson, P. & Harryson, P. Structural investigation of disordered stress proteins. Comparison of full-length dehydrins with isolated peptides of their conserved segments. *Plant Physiol.* **141**, 638–650 (2006).

253. Kovacs, D., Kalmar, E., Torok, Z. & Tompa, P. Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. *Plant Physiol.* **147**, 381–390 (2008).

254. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–261 (2004).

255. Schlicker, A., Domingues, F. S., Rahnenführer, J. & Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**, 302 (2006).

256. Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* **7**, e1002073 (2011).

257. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208 (2006).

258. Buljan, M. *et al.* Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell* **46**, 871–883 (2012).

259. Khatri, P. & Drăghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**, 3587–3595 (2005).

260. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).

261. Yates, F. Contingency Tables Involving Small Numbers and the χ2 Test. *Supplement to the Journal of the Royal Statistical Society* **1**, 217–235 (1934).

262. Benjamini, Y. & Y, H. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**, 289 – 300 (1995).

263. R Development Core Team *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2009).at <http://www.R-project.org>

264. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE* **6**, e21800 (2011).

265. Wang, H. & Deng, X. W. Arabidopsis FHY3 defines a key phytochrome A signaling component directly interacting with its homologous partner FAR1. *EMBO J.* **21**, 1339–1349 (2002).

266. Chen, M. & Chory, J. Phytochrome signaling mechanisms and the control of plant development. *Trends Cell Biol.* **21**, 664–671 (2011).

267. Bae, G. & Choi, G. Decoding of light signals by plant phytochromes and their interacting proteins. *Annu Rev Plant Biol* **59**, 281–311 (2008).

268. Franklin, K. A. & Quail, P. H. Phytochrome functions in Arabidopsis development. *J. Exp. Bot.* **61**, 11–24 (2010).

269. Goossens, A. *et al.* A functional genomics approach toward the understanding of secondary metabolism in plant cells. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8595–8600 (2003).

270. Rhodes, M. J. C. Physiological roles for secondary metabolites in plants: some progress, many outstanding problems. *Plant Molecular Biology* **24**, 1–20 (1994).

271. Seigler, D. S. *Plant Secondary Metabolism*. (Springer, 1998).

272. Zhuo, Y. *et al.* Dynamic interactions between clathrin and locally structured elements in a disordered protein mediate clathrin lattice assembly. *J. Mol. Biol.* **404**, 274–290 (2010).

273. Scheele, U. *et al.* Molecular and functional characterization of clathrin- and AP-2-binding determinants within a disordered domain of auxilin. *J. Biol. Chem.* **278**, 25357–25368 (2003).

274. Dafforn, T. R. & Smith, C. J. I. Natively unfolded domains in endocytosis: hooks, lines and linkers. *EMBO Reports* **5**, 1046–1052 (2004).

275. Kalthoff, C., Alves, J., Urbanke, C., Knorr, R. & Ungewickell, E. J. Unusual structural organization of the endocytic proteins AP180 and epsin 1. *The Journal of biological chemistry* **277**, 8209–8216 (2002).

276. Stagg, S. M., LaPointe, P. & Balch, W. E. Structural design of cage and coat scaffolds that direct membrane traffic. *Current opinion in structural biology* **17**, 221–228 (2007).

277. Spang, A. The life cycle of a transport vesicle. *Cell. Mol. Life Sci.* **65**, 2781–2789 (2008).

278. Reider, A. & Wendland, B. Endocytic adaptors–social networking at the plasma membrane. *Journal of cell science* **124**, 1613–1622 (2011).

279. Evans, P. R. & Owen, D. J. Endocytosis and vesicle trafficking. *Curr. Opin. Struct. Biol.* **12**, 814–821 (2002).

280. Jahn, R. & Scheller, R. H. SNAREs–engines for membrane fusion. *Nature reviews. Molecular cell biology* **7**, 631–643 (2006).

281. Südhof, T. C. & Rothman, J. E. Membrane fusion: grappling with SNARE and SM proteins. *Science (New York, N.Y.)* **323**, 474–477 (2009).

282. Bröcker, C., Engelbrecht-Vandré, S. & Ungermann, C. Multisubunit tethering complexes and their role in membrane fusion. *Current biology: CB* **20**, R943–952 (2010).

283. Malsam, J., Kreye, S. & Söllner, T. H. Membrane fusion: SNAREs and regulation. *Cellular and molecular life sciences: CMLS* **65**, 2814–2832 (2008).

284. Cai, H., Reinisch, K. & Ferro-Novick, S. Coats, tethers, Rabs, and SNAREs work together to mediate the intracellular destination of a transport vesicle. *Developmental cell* **12**, 671–682 (2007).

285. Hsu, V. W. & Yang, J.-S. Mechanisms of COPI vesicle formation. *FEBS letters* **583**, 3758–3763 (2009).

286. Sato, K. & Nakano, A. Mechanisms of COPII vesicle formation and protein sorting. *FEBS letters* **581**, 2076–2082 (2007).

287. González-Gaitán, M. & Jäckle, H. Role of Drosophila alpha-adaptin in presynaptic vesicle recycling. *Cell* **88**, 767–776 (1997).

288. Yeung, B. G., Phan, H. L. & Payne, G. S. Adaptor complex-independent clathrin function in yeast. *Molecular biology of the cell* **10**, 3643–3659 (1999).

289. Payne, G. S. & Schekman, R. A test of clathrin function in protein secretion and cell growth. *Science (New York, N.Y.)* **230**, 1009–1014 (1985).

290. Wendland, B., Steece, K. E. & Emr, S. D. Yeast epsins contain an essential N-terminal ENTH domain, bind clathrin and are required for endocytosis. *The EMBO journal* **18**, 4383–4393 (1999).

291. Wesp, A. *et al.* End4p/Sla2p interacts with actin-associated proteins for endocytosis in Saccharomyces cerevisiae. *Molecular biology of the cell* **8**, 2291–2306 (1997).

292. Allen, C. L., Goulding, D. & Field, M. C. Clathrin-mediated endocytosis is essential in Trypanosoma brucei. *The EMBO journal* **22**, 4991–5002 (2003).

293. Baines, A. C. & Zhang, B. Receptor-mediated protein transport in the early secretory pathway. *Trends in biochemical sciences* **32**, 381–388 (2007).

294. Moelleken, J. *et al.* Differential localization of coatomer complex isoforms within the Golgi apparatus. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 4425–4430 (2007).

295. Pagano, A. *et al.* Sec24 proteins and sorting at the endoplasmic reticulum. *The Journal of biological chemistry* **274**, 7833–7840 (1999).

296. Robinson, M. S. Adaptable adaptors for coated vesicles. *Trends in Cell Biology* **14**, 167–174 (2004).

297. Reider, A. & Wendland, B. Endocytic adaptors--social networking at the plasma membrane. *J. Cell. Sci.* **124**, 1613–1622 (2011).

298. Fuxreiter, M., Tompa, P. & Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics (Oxford, England)* **23**, 950–956 (2007).

299. Davey, N. E., Edwards, R. J. & Shields, D. C. Computational identification and analysis of protein short linear motifs. *Frontiers in bioscience: a journal and virtual library* **15**, 801–825 (2010).

300. Eisenhaber, B. & Eisenhaber, F. Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? *Current protein & peptide science* **8**, 197–203 (2007).

301. Weatheritt, R. J. & Gibson, T. J. Linear motifs: lost in (pre)translation. *Trends in biochemical sciences* **37**, 333–341 (2012).

302. Hegyi, H., Schad, E. & Tompa, P. Structural disorder promotes assembly of protein complexes. *BMC structural biology* **7**, 65 (2007).

303. Sannerud, R., Saraste, J. & Goud, B. Retrograde traffic in the biosynthetic-secretory route: pathways and machinery. *Current opinion in cell biology* **15**, 438–445 (2003).

304. Doherty, G. J. & McMahon, H. T. Mechanisms of endocytosis. *Annual review of biochemistry* **78**, 857–902 (2009).

305. Gerst, J. E. SNARE regulators: matchmakers and matchbreakers. *Biochimica et biophysica acta* **1641**, 99–110 (2003).

306. Consortium, U. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* **40**, D71–75 (2012).

307. Punta, M. *et al.* The Pfam protein families database. *Nucleic acids research* **40**, D290–301 (2012).

308. Ostlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research* **38**, D196–D203 (2009).

309. Sussman, J. L. *et al.* Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta crystallographica. Section D, Biological crystallography* **54**, 1078–1084 (1998).

310. Jahn, R. Sec1/Munc18 Proteins: Mediators of Membrane Fusion Moving to Center Stage. *Neuron* **27**, 201–204 (2000).

311. Bracher, A. & Weissenhorn, W. Structural basis for the Golgi membrane recruitment of Sly1p by Sed5p. *EMBO J.* **21**, 6114–6124 (2002).

312. Hu, S.-H., Latham, C. F., Gee, C. L., James, D. E. & Martin, J. L. Structure of the Munc18c/Syntaxin4 N-peptide complex defines universal features of the N-peptide binding mode of Sec1/Munc18 proteins. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8773–8778 (2007).

313. Burkhardt, P., Hattendorf, D. A., Weis, W. I. & Fasshauer, D. Munc18a controls SNARE assembly through its interaction with the syntaxin N-peptide. *EMBO J.* **27**, 923–933 (2008).

314. Brett, T. J., Traub, L. M. & Fremont, D. H. Accessory protein recruitment motifs in clathrin-mediated endocytosis. *Structure (London, England: 1993)* **10**, 797–809 (2002).

315. Dyson, H. J. Expanding the proteome: disordered and alternatively folded proteins. *Quarterly reviews of biophysics* **44**, 467–518 (2011).

316. von Ossowski, I. *et al.* Protein disorder: conformational distribution of the flexible linker in a chimeric double cellulase. *Biophysical journal* **88**, 2823–2832 (2005).

317. Tompa, P. On the supertertiary structure of proteins. *Nature chemical biology* **8**, 597–600 (2012).

318. Pancsa, R. & Fuxreiter, M. Interactions via intrinsically disordered regions: What kind of motifs? *IUBMB Life* **64**, 513–520 (2012).

319. Gürkan, C., Stagg, S. M., Lapointe, P. & Balch, W. E. The COPII cage: unifying principles of vesicle coat assembly. *Nature reviews. Molecular cell biology* **7**, 727–738 (2006).

320. Kümmel, D. *et al.* Complexin cross-links prefusion SNAREs into a zigzag array. *Nature structural & molecular biology* **18**, 927–933 (2011).

321. Chen, X. *et al.* Three-dimensional structure of the complexin/SNARE complex. *Neuron* **33**, 397–409 (2002).

322. Giraudo, C. G., Eng, W. S., Melia, T. J. & Rothman, J. E. A clamping mechanism involved in SNARE-dependent exocytosis. *Science* **313**, 676–680 (2006).

323. Schaub, J. R., Lu, X., Doneske, B., Shin, Y.-K. & McNew, J. A. Hemifusion arrest by complexin is relieved by Ca2+-synaptotagmin I. *Nature structural & molecular biology* **13**, 748–750 (2006).

324. Ngô, H. M. *et al.* AP-1 in Toxoplasma gondii mediates biogenesis of the rhoptry secretory organelle from a post-Golgi compartment. *The Journal of biological chemistry* **278**, 5343–5352 (2003).

325. Lefkir, Y. *et al.* Involvement of the AP-1 adaptor complex in early steps of phagocytosis and macropinocytosis. *Molecular biology of the cell* **15**, 861–869 (2004).

326. Dwyer, N. D., Adler, C. E., Crump, J. G., L'Etoile, N. D. & Bargmann, C. I. Polarized dendritic transport and the AP-1 mu1 clathrin adaptor UNC-101 localize odorant receptors to olfactory cilia. *Neuron* **31**, 277–287 (2001).

327. Fölsch, H., Ohno, H., Bonifacino, J. S. & Mellman, I. A novel clathrin adaptor complex mediates basolateral targeting in polarized epithelial cells. *Cell* **99**, 189–198 (1999).

328. Murthy, V. N. & De Camilli, P. Cell biology of the presynaptic terminal. *Annual review of neuroscience* **26**, 701–728 (2003).

329. Dell'Angelica, E. C., Mullins, C., Caplan, S. & Bonifacino, J. S. Lysosome-related organelles. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **14**, 1265–1278 (2000).

330. Berdnik, D., Török, T., González-Gaitán, M. & Knoblich, J. A. The endocytic protein alpha-Adaptin is required for numb-mediated asymmetric cell division in Drosophila. *Developmental cell* **3**, 221–231 (2002).

331. Mészáros, B., Simon, I. & Dosztányi, Z. Prediction of protein binding regions in disordered proteins. *PLoS computational biology* **5**, e1000376 (2009).

332. Fernández, A. & Scott, R. Dehydron: a structurally encoded signal for protein interaction. *Biophys. J.* **85**, 1914–1928 (2003).

333. Fernández, A. Keeping dry and crossing membranes. *Nat. Biotechnol.* **22**, 1081–1084 (2004).

334. Fernández, A. & Scheraga, H. A. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 113–118 (2003).

335. Fernández, A. & Scott, L. R. Adherence of Packing Defects in Soluble Proteins. *Phys. Rev. Lett.* **91**, 018102 (2003).

336. Fernández, A. & Berry, R. S. Molecular dimension explored in evolution to promote proteomic complexity. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13460–13465 (2004).

337. Fernández, A., Sosnick, T. R. & Colubri, A. Dynamics of Hydrogen Bond Desolvation in Protein Folding. *Journal of Molecular Biology* **321**, 659–675 (2002).

338. Fernández, A., Kardos, J., Scott, L. R., Goto, Y. & Berry, R. S. Structural defects and the diagnosis of amyloidogenic propensity. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 6446–6451 (2003).

339. Fernández, A. & Berry, R. S. Proteins with H-bond packing defects are highly interactive with lipid bilayers: Implications for amyloidogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 2391–2396 (2003).

340. Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5772–5777 (2003).

341. Ma, B. *et al.* Comparison of the protein–protein interfaces in the p53–DNA crystal structures: Towards elucidation of the biological interface. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3988–3993 (2005).

342. Rajamani, D., Thiel, S., Vajda, S. & Camacho, C. J. Anchor residues in protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11287–11292 (2004).

343. Vacic, V. *et al.* Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* **6**, 2351–2366 (2007).

344. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701–1718 (2005).

345. Rizzo, R. C. & Jorgensen, W. L. OPLS All-Atom Model for Amines: Resolution of the Amine Hydration Problem. *J. Am. Chem. Soc.* **121**, 4827–4836 (1999).

346. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935 (1983).

347. Ho, W. C., Fitzgerald, M. X. & Marmorstein, R. Structure of the p53 core domain dimer bound to DNA. *J. Biol. Chem.* **281**, 20494–20502 (2006).

348. Fernández, A. What caliber pore is like a pipe? Nanotubes as modulators of ionic gradients. *The Journal of Chemical Physics* **119**, 5315–5319 (2003).

349. Meador, W. E., Means, A. R. & Quiocho, F. A. Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures. *Science* **262**, 1718–1721 (1993).

350. Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. Atomic structure of the actin:DNase I complex. *Nature* **347**, 37–44 (1990).

351. Williams, R. S., Green, R. & Glover, J. N. Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1. *Nat. Struct. Biol.* **8**, 838–842 (2001).

352. Zahn, R. *et al.* NMR solution structure of the human prion protein. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 145–150 (2000).

353. Schnuchel, A., Wiltscheck, R., Eichinger, L., Schleicher, M. & Holak, T. A. Structure of severin domain 2 in solution. *J. Mol. Biol.* **247**, 21–27 (1995).

354. Glover, J. N. & Harrison, S. C. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373**, 257–261 (1995).

355. Lavigne, P. *et al.* Insights into the mechanism of heterodimerization from the 1H-NMR solution structure of the c-Myc-Max heterodimeric leucine zipper. *J. Mol. Biol.* **281**, 165–181 (1998).

356. Han, J.-D. J. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).

357. González-Ruiz, D. & Gohlke, H. Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr. Med. Chem.* **13**, 2607–2625 (2006).

358. Mitraki, A. Protein aggregation from inclusion bodies to amyloid and biomaterials. *Adv Protein Chem Struct Biol* **79**, 89–125 (2010).

359. Silverman, R. B. The potential use of mechanism-based enzyme inactivators in medicine. *J. Enzym. Inhib.* **2**, 73–90 (1988).

360. Reynolds, P. E. Control of peptidoglycan synthesis in vancomycin-resistant enterococci: D,D-peptidases and D,D-carboxypeptidases. *Cell. Mol. Life Sci.* **54**, 325–331 (1998).

361. Voss, K. & Galensa, R. Determination of L- and D-amino acids in foodstuffs by coupling of high-performance liquid chromatography with enzyme reactors. *Amino Acids* **18**, 339–352 (2000).

362. Gandolfi, I. *et al.* D-alanine in Fruit Juices: A Molecular Marker of Bacterial Activity, Heat Treatments and Shelf-life. *Journal of Food Science* **59**, 152–154 (1994).

363. Pietrosemoli, N., Crespo, A. & Fernandez, A. Dehydration Propensity of Order-Disorder Intermediate Regions in Soluble Proteins. *Journal of Proteome Research* **6**, 3519–3526 (2007).

Appendix A

Table 1A. GO:BP terms enriched in disorder in *A. thaliana.*

term ID	description	frequency	uniqueness
GO:0000003	reproduction	2.31%	1
GO:0007389	pattern specification process	0.07%	0.78
GO:0007623	circadian rhythm	0.02%	0.99
GO:0009058	biosynthetic process	31.17%	0.97
GO:0009639	response to red or far red light	0.01%	0.86
GO:0009266	response to temperature stimulus	0.14%	0.84
GO:0009314	response to radiation	0.19%	0.84
GO:0009416	response to light stimulus	0.17%	0.84
GO:0009987	cellular process	68.00%	1
GO:0016043	cellular component organization	3.68%	0.9
GO:0006323	DNA packaging	0.32%	0.79
GO:0034728	nucleosome organization	0.19%	0.87
GO:0016568	chromatin modification	0.11%	0.87
GO:0051276	chromosome organization	0.43%	0.87
GO:0006333	chromatin assembly or disassembly	0.20%	0.87
GO:0006325	chromatin organization	0.31%	0.87
GO:0006996	organelle organization	0.84%	0.87
GO:0016192	vesicle-mediated transport	0.35%	0.96
GO:0022406	membrane docking	0.03%	0.97
GO:0022414	reproductive process	2.25%	0.94
GO:0003006	developmental process involved in reproduction	0.14%	0.79
GO:0030005	cellular di-, tri-valent inorganic cation homeostasis	0.15%	0.99
GO:0032501	multicellular organismal process	1.47%	0.99
GO:0032502	developmental process	1.85%	0.99
GO:0051179	localization	19.12%	0.99

GO:0055080	cation homeostasis	0.26%	0.83
GO:0065007	biological regulation	15.11%	0.99
GO:0043687	post-translational protein modification	0.01%	0.91
GO:0022402	cell cycle process	0.22%	0.96
GO:0051301	cell division	1.07%	0.96
GO:0007049	cell cycle	1.21%	0.96
GO:0043170	macromolecule metabolic process	35.19%	0.97
GO:0006807	nitrogen compound metabolic process	36.48%	0.97
GO:0044237	cellular metabolic process	56.72%	0.93
GO:0016310	phosphorylation	6.14%	0.91
GO:0006796	phosphate-containing compound metabolic process	6.73%	0.91
GO:0006793	phosphorus metabolic process	6.76%	0.93
GO:0009250	glucan biosynthetic process	0.15%	0.89
GO:0008380	RNA splicing	0.17%	0.87
GO:0043412	macromolecule modification	5.67%	0.9
GO:0006139	nucleobase-containing compound metabolic process	29.19%	0.87
GO:0044249	cellular biosynthetic process	29.57%	0.87
GO:0032940	secretion by cell	0.61%	0.92
GO:0048278	vesicle docking	0.03%	0.93
GO:0006887	exocytosis	0.05%	0.92
GO:0006904	vesicle docking involved in exocytosis	0.02%	0.92
GO:0010646	regulation of cell communication	0.18%	0.8
GO:0009719	response to endogenous stimulus	0.22%	0.86
GO:0046903	secretion	0.63%	0.96
GO:0048518	positive regulation of biological process	0.52%	0.81
GO:0030001	metal ion transport	1.43%	0.95
GO:0051173	positive regulation of nitrogen compound metabolic process	0.28%	0.7
GO:0048522	positive regulation of cellular process	0.47%	0.72
GO:0031328	positive regulation of cellular biosynthetic process	0.29%	0.69
GO:0031325	positive regulation of cellular metabolic process	0.33%	0.7

GO:0045935	positive regulation of nucleobase-containing compound metabolic process	0.27%	0.68
GO:0010628	positive regulation of gene expression	0.21%	0.69
GO:0045893	positive regulation of transcription, DNA-dependent	0.20%	0.66
GO:0009893	positive regulation of metabolic process	0.34%	0.72
GO:0010604	positive regulation of macromolecule metabolic process	0.32%	0.69
GO:0009891	positive regulation of biosynthetic process	0.29%	0.71
GO:0010557	positive regulation of macromolecule biosynthetic process	0.23%	0.69
GO:0010467	gene expression	17.65%	0.88
GO:0016071	mRNA metabolic process	0.72%	0.86
GO:0006457	protein folding	0.97%	0.88
GO:0009628	response to abiotic stimulus	0.40%	0.85
GO:0006259	DNA metabolic process	7.22%	0.83
GO:0010033	response to organic substance	0.36%	0.81
GO:0042221	response to chemical stimulus	1.88%	0.83
GO:0006810	transport	18.62%	0.94
GO:0051234	establishment of localization	18.63%	0.94
GO:0044260	cellular macromolecule metabolic process	30.87%	0.84
GO:0010200	response to chitin	0.00%	0.83
GO:0006974	response to DNA damage stimulus	1.94%	0.79
GO:0006284	base-excision repair	0.23%	0.75
GO:0006281	DNA repair	1.92%	0.69
GO:0080090	regulation of primary metabolic process	9.23%	0.68
GO:0010468	regulation of gene expression	8.98%	0.64
GO:0031323	regulation of cellular metabolic process	9.22%	0.67
GO:0031326	regulation of cellular biosynthetic process	8.77%	0.65
GO:0051252	regulation of RNA metabolic process	8.55%	0.6
GO:0019219	regulation of nucleobase-containing compound metabolic process	8.85%	0.64
GO:0006355	regulation of transcription, DNA-dependent	8.53%	0.59
GO:0060255	regulation of macromolecule metabolic process	9.28%	0.64
GO:0009889	regulation of biosynthetic process	8.78%	0.68

GO:0010556	regulation of macromolecule biosynthetic process	8.75%	0.63
GO:0051171	regulation of nitrogen compound metabolic process	8.86%	0.67
GO:0009966	regulation of signal transduction	0.44%	0.68
GO:0051056	regulation of small GTPase mediated signal transduction	0.26%	0.69
GO:0046578	regulation of Ras protein signal transduction	0.21%	0.7
GO:0006464	protein modification process	4.00%	0.85
GO:0006468	protein phosphorylation	2.19%	0.84
GO:0016070	RNA metabolic process	13.96%	0.81
GO:0016567	protein ubiquitination	0.11%	0.89
GO:0034645	cellular macromolecule biosynthetic process	19.25%	0.8
GO:0009059	macromolecule biosynthetic process	19.47%	0.85
GO:0034641	cellular nitrogen compound metabolic process	35.14%	0.87
GO:0009751	response to salicylic acid stimulus	0.01%	0.82
GO:0048580	regulation of post-embryonic development	0.01%	0.68
GO:0009739	response to gibberellin stimulus	0.01%	0.82
GO:0048367	shoot development	0.02%	0.77
GO:0048827	phyllome development	0.01%	0.78
GO:0048366	leaf development	0.01%	0.78
GO:0009887	organ morphogenesis	0.10%	0.76
GO:0009965	leaf morphogenesis	0.00%	0.79
GO:0010016	shoot morphogenesis	0.01%	0.78
GO:0016044	cellular membrane organization	0.17%	0.9
GO:0022621	shoot system development	0.02%	0.79
GO:0065004	protein-DNA complex assembly	0.19%	0.88
GO:0006334	nucleosome assembly	0.19%	0.78
GO:0030154	cell differentiation	0.38%	0.76
GO:0048513	organ development	0.34%	0.75
GO:0048856	anatomical structure development	1.48%	0.76
GO:0048869	cellular developmental process	1.13%	0.75
GO:0050793	regulation of developmental process	0.85%	0.61

GO:0007275	multicellular organismal development	0.90%	0.76
GO:0048468	cell development	0.18%	0.77
GO:0009653	anatomical structure morphogenesis	1.07%	0.76
GO:0048731	system development	0.55%	0.75
GO:0051239	regulation of multicellular organismal process	0.21%	0.7
GO:0000160	two-component signal transduction system (phosphorelay)	2.35%	0.64
GO:0035556	intracellular signal transduction	2.60%	0.63
GO:0007165	signal transduction	5.49%	0.6
GO:0009873	ethylene mediated signaling pathway	0.00%	0.71
GO:0006397	mRNA processing	0.62%	0.85
GO:0051716	cellular response to stimulus	7.64%	0.79
GO:0009743	response to carbohydrate stimulus	0.03%	0.81
GO:0009755	hormone-mediated signaling pathway	0.11%	0.65
GO:0009753	response to jasmonic acid stimulus	0.01%	0.81
GO:0009737	response to abscisic acid stimulus	0.03%	0.8
GO:0009734	auxin mediated signaling pathway	0.02%	0.69
GO:0009733	response to auxin stimulus	0.02%	0.8
GO:0009725	response to hormone stimulus	0.19%	0.78
GO:0009723	response to ethylene stimulus	0.01%	0.81
GO:0032870	cellular response to hormone stimulus	0.13%	0.76
GO:0009791	post-embryonic development	0.10%	0.78
GO:0006351	transcription, DNA-dependent	10.06%	0.77
GO:0048608	reproductive structure development	0.08%	0.78
GO:0009908	flower development	0.02%	0.77
GO:0048467	gynoecium development	0.00%	0.78
GO:0048440	carpel development	0.00%	0.78
GO:0048438	floral whorl development	0.01%	0.78
GO:0019222	regulation of metabolic process	9.78%	0.73
GO:0050794	regulation of cellular process	14.07%	0.7
GO:0050789	regulation of biological process	14.66%	0.73

GO:0009790	embryo development	0.16%	0.77

Table 2A. GO:BP terms enriched in disorder in *A. thaliana* with respect to *H. sapiens.*

term ID	description	frequency	uniqueness
GO:0006826	iron ion transport	0.19%	0.96
GO:0009581	detection of external stimulus	0.11%	0.93
GO:0019748	secondary metabolic process	0.08%	0.96
GO:0055072	iron ion homeostasis	0.17%	0.85
GO:0065007	biological regulation	15.11%	0.98
GO:0009812	flavonoid metabolic process	0.01%	0.93
GO:0046434	organophosphate catabolic process	0.02%	0.95
GO:0042440	pigment metabolic process	0.32%	0.96
GO:0071941	nitrogen cycle metabolic process	0.17%	0.88
GO:0006662	glycerol ether metabolic process	0.18%	0.83
GO:0006639	acylglycerol metabolic process	0.02%	0.74
GO:0006641	triglyceride metabolic process	0.01%	0.76
GO:0006807	nitrogen compound metabolic process	36.48%	0.95
GO:0006081	cellular aldehyde metabolic process	0.12%	0.92
GO:0006638	neutral lipid metabolic process	0.02%	0.84
GO:0051186	cofactor metabolic process	3.54%	0.89
GO:0006457	protein folding	0.97%	0.83
GO:0022900	electron transport chain	4.73%	0.88
GO:0042726	flavin-containing compound metabolic process	0.22%	0.79
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	0.04%	0.8
GO:0000375	RNA splicing, via transesterification reactions	0.04%	0.8
GO:0006139	nucleobase-containing compound metabolic process	29.19%	0.74

GO:0010467	gene expression	17.65%	0.86
GO:0018904	organic ether metabolic process	0.18%	0.86
GO:0019400	alditol metabolic process	0.31%	0.81
GO:0032940	secretion by cell	0.61%	0.91
GO:0006887	exocytosis	0.05%	0.92
GO:0009410	response to xenobiotic stimulus	0.02%	0.92
GO:0046903	secretion	0.63%	0.96
GO:0042364	water-soluble vitamin biosynthetic process	1.15%	0.72
GO:0006520	cellular amino acid metabolic process	5.56%	0.64
GO:0008652	cellular amino acid biosynthetic process	3.17%	0.58
GO:0016053	organic acid biosynthetic process	3.99%	0.66
GO:0043436	oxoacid metabolic process	7.03%	0.71
GO:0046394	carboxylic acid biosynthetic process	3.97%	0.66
GO:0019752	carboxylic acid metabolic process	7.03%	0.71
GO:0009067	aspartate family amino acid biosynthetic process	0.69%	0.64
GO:0042742	defense response to bacterium	0.09%	0.92
GO:0009617	response to bacterium	0.12%	0.92
GO:0006183	GTP biosynthetic process	0.04%	0.7
GO:0046131	pyrimidine ribonucleoside metabolic process	0.18%	0.67
GO:0009220	pyrimidine ribonucleotide biosynthetic process	0.17%	0.63
GO:0072528	pyrimidine-containing compound biosynthetic process	0.58%	0.65
GO:0046051	UTP metabolic process	0.04%	0.69
GO:0006220	pyrimidine nucleotide metabolic process	0.62%	0.64
GO:0006228	UTP biosynthetic process	0.04%	0.65
GO:0006221	pyrimidine nucleotide biosynthetic process	0.55%	0.6
GO:0006778	porphyrin-containing compound metabolic process	0.70%	0.74
GO:0006779	porphyrin-containing compound biosynthetic process	0.68%	0.67
GO:0033014	tetrapyrrole biosynthetic process	0.76%	0.67

GO:0042168	heme metabolic process	0.10%	0.78
GO:0006082	organic acid metabolic process	7.12%	0.75
GO:0009059	macromolecule biosynthetic process	19.47%	0.77
GO:0033013	tetrapyrrole metabolic process	0.77%	0.77
GO:0009309	amine biosynthetic process	3.26%	0.69
GO:0006414	translational elongation	0.67%	0.78
GO:0046471	phosphatidylglycerol metabolic process	0.00%	0.84
GO:0006289	nucleotide-excision repair	0.23%	0.75
GO:0000160	two-component signal transduction system (phosphorelay)	2.35%	0.71
GO:0090304	nucleic acid metabolic process	21.07%	0.73
GO:0034645	cellular macromolecule biosynthetic process	19.25%	0.69
GO:0016070	RNA metabolic process	13.96%	0.68
GO:0006720	isoprenoid metabolic process	0.40%	0.81
GO:0019222	regulation of metabolic process	9.78%	0.73
GO:0050794	regulation of cellular process	14.07%	0.7
GO:0050789	regulation of biological process	14.66%	0.74
GO:0070887	cellular response to chemical stimulus	0.37%	0.88
GO:0034641	cellular nitrogen compound metabolic process	35.14%	0.75
GO:0006721	terpenoid metabolic process	0.23%	0.79
GO:0016108	tetraterpenoid metabolic process	0.04%	0.81
GO:0016116	carotenoid metabolic process	0.04%	0.81
GO:0043288	apocarotenoid metabolic process	0.00%	0.83
GO:0071310	cellular response to organic substance	0.23%	0.88
GO:0006351	transcription, DNA-dependent	10.06%	0.63
GO:0032774	RNA biosynthetic process	10.14%	0.65
GO:0016101	diterpenoid metabolic process	0.01%	0.83
GO:0080090	regulation of primary metabolic process	9.23%	0.65
GO:0009889	regulation of biosynthetic process	8.78%	0.62

GO:0010468	regulation of gene expression	8.98%	0.63
GO:0031323	regulation of cellular metabolic process	9.22%	0.64
GO:0031326	regulation of cellular biosynthetic process	8.77%	0.57
GO:0051252	regulation of RNA metabolic process	8.55%	0.52
GO:0010556	regulation of macromolecule biosynthetic process	8.75%	0.58
GO:0019219	regulation of nucleobase-containing compound metabolic process	8.85%	0.56
GO:2000112	regulation of cellular macromolecule biosynthetic process	8.75%	0.54
GO:0006355	regulation of transcription, DNA-dependent	8.53%	0.49
GO:0060255	regulation of macromolecule metabolic process	9.28%	0.65
GO:0051171	regulation of nitrogen compound metabolic process	8.86%	0.6

Table 3A. Summary of intrinsic disorder metrics for A. thaliana and H. sapiens for the different prediction methods.

	Mean content of disorder (%)		Proteins with at least one LDW (%)		Mean number of LDW		Mean number of residues belonging to LDWs	
	A.th.	*H.sa.*	*A.th.*	*H.sa.*	*A.th.*	*H.sa.*	*A.th.*	*H.sa.*
IuPred (short)	17.0	22.4	36.3	60.9	0.6	1.4	9.0	14.9
IuPred (long)	16.8	24.5	32.8	56.6	0.6	1.3	10.5	17.6
VSL2	38.9	44.9	68.1	78.9	1.2	1.9	26.1	34.4

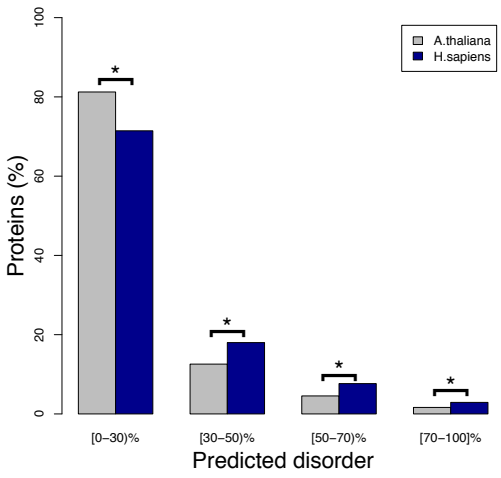

Figure 1A. Fraction of proteins with different degrees of predicted disorder in *A. thaliana and H. sapiens*. Protein disorder (as the percentage of disordered residues with respect to the sequence length) is binned into different ranges. Data based on IuPred (short) predictions.

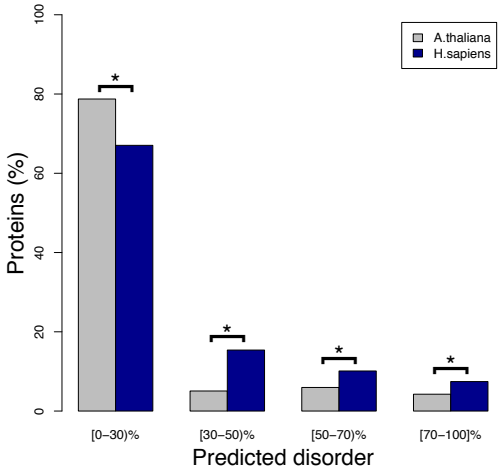

Figure 2A. Fraction of proteins with different degrees of predicted disorder in *A. thaliana and H. sapiens*. Protein disorder (as the percentage of disordered residues with respect to the sequence length) is binned into different ranges. Data based on IuPred (long) predictions.

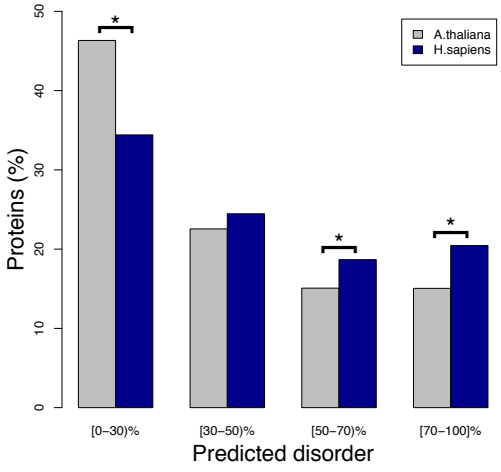

Figure 3A. Fraction of proteins with different degrees of predicted disorder in *A. thaliana* and *H. sapiens*. Protein disorder (as the percentage of disordered residues with respect to the sequence length) is binned into different ranges. Data based on VSL2 predictions.

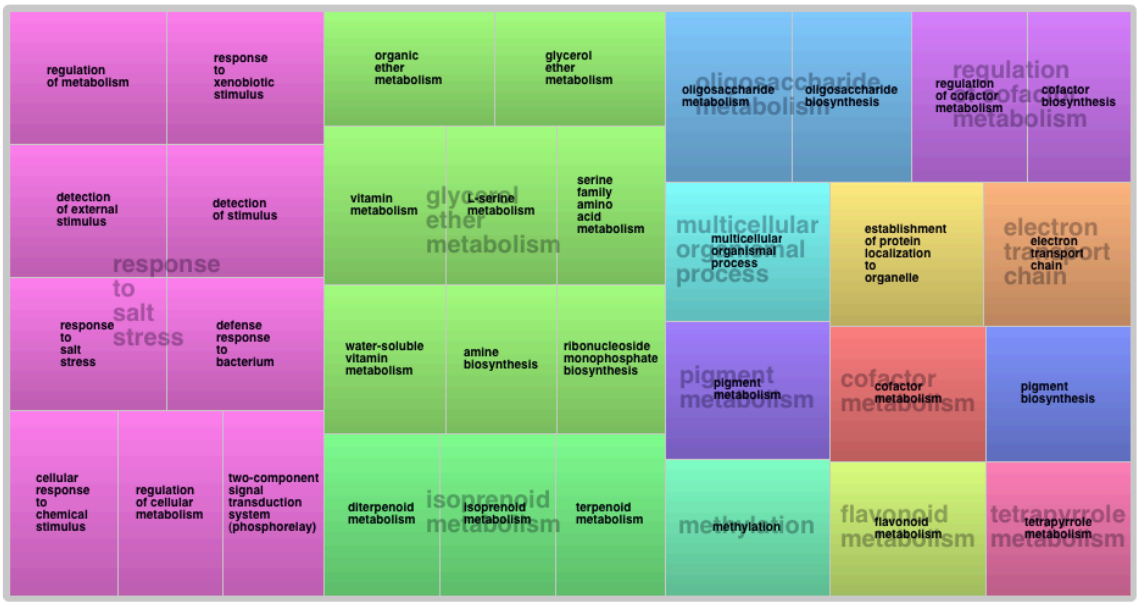

Figure 4A. Representation of the main GO "Biological Processes" comparatively enriched in disordered proteins in *A. thaliana* respect to *H. sapiens.* Disordered proteins correspond to those with 1 or more LDWs based on VSL2 predictions. Figure adapted from REVIGO, a method for summarizing and visualizing lists of GO terms. Each rectangle represents a cluster of related terms labelled according to a representative term. Rectangles are grouped in "superclusters" (identified with the same color) based on SimRel semantic similarity measure.

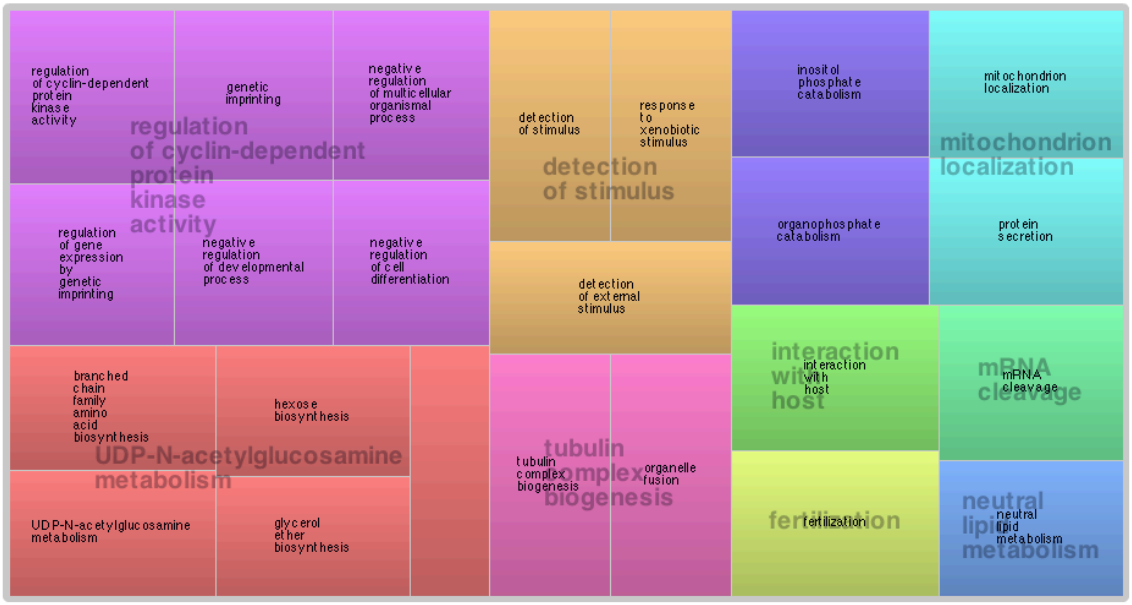

Figure 5A. Representation of the GO "Biological Processes" comparatively enriched in disordered proteins in *A. thaliana* respect to *H. sapiens.* Disordered proteins correspond to those with 1 or more LDWs based on Iupred (option "long") predictions. Same REVIGO representation adaptation as in Figure 4A.

Figure 6A. Representation of the GO "Biological Processes" comparatively enriched in disordered proteins in *A. thaliana* respect to *H. sapiens.* Disordered proteins correspond to those with 1 or more LDWs based on Iupred (option "short") predictions. Same REVIGO representation adaptation as in Figure 4A.

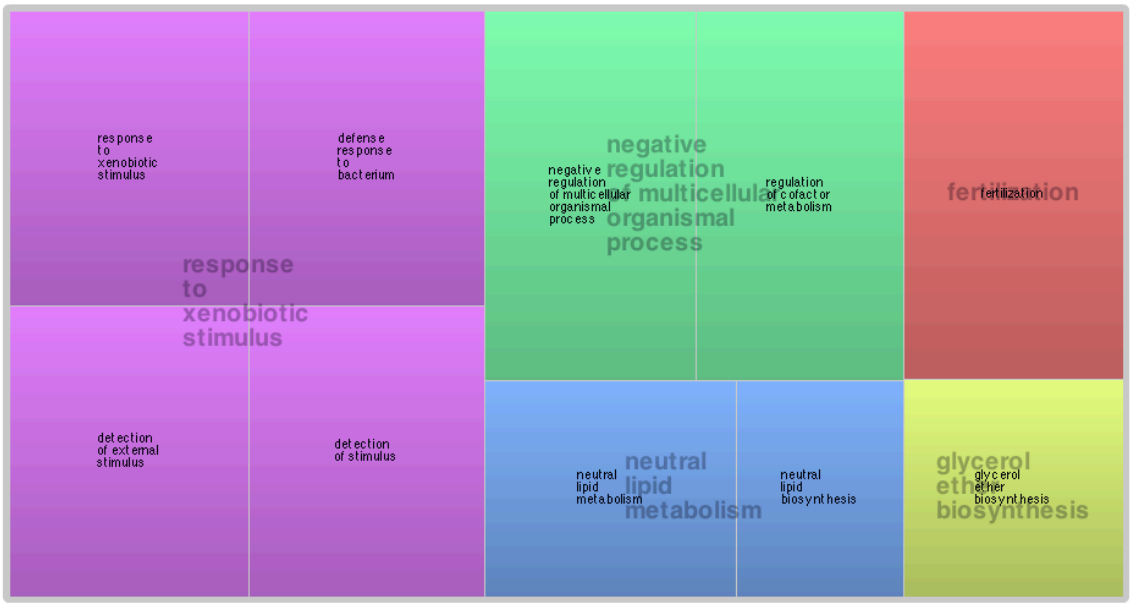

Figure 7A. Representation of the GO "Biological Processes" comparatively enriched in residues belonging to disordered binding regions (DBR) in *A. thaliana* with respect to *Human*. DBR are calculated based on ANCHOR predictions. Same REVIGO representation adaptation as in Figure 4A.

Appendix B

Table 1B. Proteins involved in the main trafficking pathways in human.

Protein name	Uniprot accession	Functional classification	System	Disordered residues	Total number of residues
AP-4 complex subunit mu-1	O00189	ADAPTOR/SORTING	CLATHRIN	29	453
AP-3 complex subunit beta-1	O00203	ADAPTOR/SORTING	CLATHRIN	254	1094
HIP1 (Huntingtin-interacting protein 1)	O00291	ADAPTOR/SORTING	CLATHRIN	221	1037
FCH domain only protein 1	O14526	ADAPTOR/SORTING	CLATHRIN	423	889
AP-3 complex subunit delta-1	O14617	ADAPTOR/SORTING	CLATHRIN	340	1153
HRS (Hepatocyte growth factor-regulated tyrosine kinase substrate)	O14964	ADAPTOR/SORTING	CLATHRIN	434	777
AP-1 complex subunit gamma-1	O43747	ADAPTOR/SORTING	CLATHRIN	85	822
Hip1R (Huntingtin-interacting protein 1-related protein)	O75146	ADAPTOR/SORTING	CLATHRIN	288	1068
DAB1	O75553	ADAPTOR/SORTING	CLATHRIN	302	588
AP-2 complex subunit alpha-2	O94973	ADAPTOR/SORTING	CLATHRIN	87	939
Epsin-2	O95208	ADAPTOR/SORTING	CLATHRIN	478	641
AP-2 complex subunit alpha-1	O95782	ADAPTOR/SORTING	CLATHRIN	144	977
Beta-arrestin-2	P32121	ADAPTOR/SORTING	CLATHRIN	64	409
Epidermal growth factor receptor substrate 15 (EPS15)	P42566	ADAPTOR/SORTING	CLATHRIN	439	896
Beta-arrestin-1	P49407	ADAPTOR/SORTING	CLATHRIN	90	418
NUMB	P49757	ADAPTOR/SORTING	CLATHRIN	399	651
AP-3 complex subunit mu-2	P53677	ADAPTOR/SORTING	CLATHRIN	19	418
AP-2 complex subunit sigma	P53680	ADAPTOR/SORTING	CLATHRIN	0	142
AP-1 complex subunit sigma-2	P56377	ADAPTOR/SORTING	CLATHRIN	0	157
AP-3 complex subunit sigma-2	P59780	ADAPTOR/SORTING	CLATHRIN	0	193

AP-1 complex subunit sigma-1A	P61966	ADAPTOR/SORTING	CLATHRIN	0	158
AP-2 complex subunit beta	P63010	ADAPTOR/SORTING	CLATHRIN	39	937
DAB2	P98082	ADAPTOR/SORTING	CLATHRIN	571	770
FCH domain only protein 2	Q0JRZ9	ADAPTOR/SORTING	CLATHRIN	280	810
AP-1 complex subunit beta-1	Q10567	ADAPTOR/SORTING	CLATHRIN	73	949
AP-3 complex subunit beta-2	Q13367	ADAPTOR/SORTING	CLATHRIN	261	1082
AP180/PICALM/SNAP91	Q13492	ADAPTOR/SORTING	CLATHRIN	184	652
epsinR(enthoprotin, CLINT1, Epsin4, Clathrin interactor 1)	Q14677	ADAPTOR/SORTING	CLATHRIN	351	625
Intersectin-1 (ITSN1)	Q15811	ADAPTOR/SORTING	CLATHRIN	489	1721
Low density lipoprotein receptor adapter protein 1 (LDLRAP1, ARH)	Q5SW96	ADAPTOR/SORTING	CLATHRIN	83	308
Adaptin ear-binding coat-associated protein 1 (NECAP1)	Q8NC96	ADAPTOR/SORTING	CLATHRIN	128	275
Stonin-2	Q8WXE9	ADAPTOR/SORTING	CLATHRIN	365	905
AP-3 complex subunit sigma-1	Q92572	ADAPTOR/SORTING	CLATHRIN	0	193
CIN85 (SH3KBP1)	Q96B97	ADAPTOR/SORTING	CLATHRIN	444	665
AP-2 complex subunit mu	Q96CW1	ADAPTOR/SORTING	CLATHRIN	21	435
AP-1 complex subunit sigma-3	Q96PC3	ADAPTOR/SORTING	CLATHRIN	0	154
SH3-containing GRB2-like protein 3-interacting protein 1 (SGIP)	Q9BQI5	ADAPTOR/SORTING	CLATHRIN	519	828
AP-1 complex subunit mu-1	Q9BXS5	ADAPTOR/SORTING	CLATHRIN	9	423
Epsin-3	Q9H201	ADAPTOR/SORTING	CLATHRIN	432	632
Adaptin ear-binding coat-associated protein 2 (NECAP2)	Q9NVZ3	ADAPTOR/SORTING	CLATHRIN	121	263
ADP-ribosylation factor-binding protein GGA3	Q9NZ52	ADAPTOR/SORTING	CLATHRIN	287	723
Intersectin-2 (ITSN2)	Q9NZM3	ADAPTOR/SORTING	CLATHRIN	364	1697
ADP-ribosylation factor-binding protein GGA2	Q9UJY4	ADAPTOR/SORTING	CLATHRIN	128	613
ADP-ribosylation factor-binding protein GGA1	Q9UJY5	ADAPTOR/SORTING	CLATHRIN	299	639
AP-4 complex subunit epsilon-1	Q9UPM8	ADAPTOR/SORTING	CLATHRIN	102	1137
AP-3 complex subunit mu-1	Q9Y2T2	ADAPTOR/SORTING	CLATHRIN	8	418
AP-4 complex subunit sigma-1	Q9Y587	ADAPTOR/SORTING	CLATHRIN	0	144

Sorting nexin-9 (SNX9)	Q9Y5X1	ADAPTOR/SORTING	CLATHRIN	163	595
AP-4 complex subunit beta-1	Q9Y6B7	ADAPTOR/SORTING	CLATHRIN	17	739
Epsin-1	Q9Y6I3	ADAPTOR/SORTING	CLATHRIN	448	576
Stonin-1	Q9Y6Q2	ADAPTOR/SORTING	CLATHRIN	153	735
AP-1 complex subunit mu-2	Q9Y6Q5	ADAPTOR/SORTING	CLATHRIN	3	423
COPD (Coatomer subunit delta)	P48444	ADAPTOR/SORTING	COPI SYSTEM	131	511
Coatomer subunit beta	P53618	ADAPTOR/SORTING	COPI SYSTEM	39	953
COPZ1 (Coatomer subunit zeta-1)	P61923	ADAPTOR/SORTING	COPI SYSTEM	0	177
COPZ2 (Coatomer subunit zeta-2)	Q9P299	ADAPTOR/SORTING	COPI SYSTEM	34	210
COPG2 (Coatomer subunit gamma-2)	Q9UBF2	ADAPTOR/SORTING	COPI SYSTEM	25	871
COPG1 (Coatomer subunit gamma-1)	Q9Y678	ADAPTOR/SORTING	COPI SYSTEM	56	874
Protein transport protein Sec24D	O94855	ADAPTOR/SORTING	COPII SYSTEM	300	1032
Protein transport protein Sec24A	O95486	ADAPTOR/SORTING	COPII SYSTEM	373	1093
Protein transport protein Sec24B	O95487	ADAPTOR/SORTING	COPII SYSTEM	441	1268
Protein transport protein Sec24C	P53992	ADAPTOR/SORTING	COPII SYSTEM	374	1094
Protein transport protein Sec23A	Q15436	ADAPTOR/SORTING	COPII SYSTEM	38	765
Protein transport protein Sec23B	Q15437	ADAPTOR/SORTING	COPII SYSTEM	36	767
Clathrin light chain A	P09496	COAT	CLATHRIN	149	248
Clathrin light chain B	P09497	COAT	CLATHRIN	171	229
Clathrin heavy chain 2	P53675	COAT	CLATHRIN	0	1640
Clathrin heavy chain 1	Q00610	COAT	CLATHRIN	18	1675
COPE (Coatomer subunit epsilon)	O14579	COAT	COPI SYSTEM	18	308
COPB2 (Coatomer subunit beta')	P35606	COAT	COPI SYSTEM	83	906
COPA (Coatomer subunit alpha)	P53621	COAT	COPI SYSTEM	113	1224

Protein transport protein Sec31A	O94979	COAT	COPII SYSTEM	410	1220
Protein SEC13 homolog	P55735	COAT	COPII SYSTEM	21	322
Protein transport protein Sec31B	Q9NQW1	COAT	COPII SYSTEM	323	1179
GAK (Cyclin-G-associated kinase)	O14976	ENZYME/ENZYME-INTERACTOR	CLATHRIN	518	1311
Synaptojanin-2 (Synaptic inositol 1,4,5-trisphosphate 5-phosphatase 2)	O15056	ENZYME/ENZYME-INTERACTOR	CLATHRIN	470	1496
Synaptojanin-1 (Synaptic inositol 1,4,5-trisphosphate 5-phosphatase 1)	O43426	ENZYME/ENZYME-INTERACTOR	CLATHRIN	531	1573
Putative tyrosine-protein phosphatase auxilin	O75061	ENZYME/ENZYME-INTERACTOR	CLATHRIN	413	913
Dynamin-2	P50570	ENZYME/ENZYME-INTERACTOR	CLATHRIN	203	870
ADP-ribosylation factor 6 (Arf6)	P62330	ENZYME/ENZYME-INTERACTOR	CLATHRIN	0	175
Dynamin-1	Q05193	ENZYME/ENZYME-INTERACTOR	CLATHRIN	193	864
Arf-GAP with coiled-coil, ANK repeat and PH domain-containing protein 1	Q15027	ENZYME/ENZYME-INTERACTOR	CLATHRIN	163	740
AAK1 (AP2-associated protein kinase 1)	Q2M2I8	ENZYME/ENZYME-INTERACTOR	CLATHRIN	557	961
Dynamin-3	Q9UQ16	ENZYME/ENZYME-INTERACTOR	CLATHRIN	194	869
ADP-ribosylation factor 1 (Arf1)/ARF1 (ADP-ribosylation factor 1)	P84077	ENZYME/ENZYME-INTERACTOR	CLATHRIN/COPI SYSTEM	0	181
RAB6A (Ras-related protein Rab-6A)	P20340	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	29	208
ARFGAP2 (ADP-ribosylation factor GTPase-activating protein 2)	Q8N6H7	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	246	521
ARFGAP1 (ADP-ribosylation factor GTPase-activating protein 1)	Q8N6T3	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	212	406
ARFGAP3 (ADP-ribosylation factor GTPase-activating protein 3)	Q9NP61	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	189	516
GTP-binding protein SAR1a	Q9NR31	ENZYME/ENZYME-INTERACTOR	COPII SYSTEM	0	198

GTP-binding protein SAR1b	Q9Y6B6	ENZYME/ENZYME-INTERACTOR	COPII SYSTEM	0	198
Cog7 (Conserved oligomeric Golgi complex subunit 7)	P83436	MULTISUBUNIT TETHERING COMPLEX	Cog	17	770
Cog2 (Conserved oligomeric Golgi complex subunit 2)	Q14746	MULTISUBUNIT TETHERING COMPLEX	Cog	63	738
Cog1 (Conserved oligomeric Golgi complex subunit 1)	Q8WTW3	MULTISUBUNIT TETHERING COMPLEX	Cog	114	980
Cog3 (Conserved oligomeric Golgi complex subunit 3)	Q96JB2	MULTISUBUNIT TETHERING COMPLEX	Cog	55	828
Cog8 (Conserved oligomeric Golgi complex subunit 8)	Q96MW5	MULTISUBUNIT TETHERING COMPLEX	Cog	51	612
Cog4 (Conserved oligomeric Golgi complex subunit 4)	Q9H9E3	MULTISUBUNIT TETHERING COMPLEX	Cog	20	785
Cog5 (Conserved oligomeric Golgi complex subunit 5)	Q9UP83	MULTISUBUNIT TETHERING COMPLEX	Cog	65	839
Cog6 (Conserved oligomeric Golgi complex subunit 6)	Q9Y2V7	MULTISUBUNIT TETHERING COMPLEX	Cog	32	657
Vps8 (Vacuolar protein sorting-associated protein 8)	Q8N3P4	MULTISUBUNIT TETHERING COMPLEX	CORVET	128	1428
Vps33A (Vacuolar protein sorting-associated protein 33A)	Q96AX1	MULTISUBUNIT TETHERING COMPLEX	CORVET	26	596
Vps16 (Vacuolar protein sorting-associated protein 16)	Q9H269	MULTISUBUNIT TETHERING COMPLEX	CORVET	18	839
Vps11/PEP5 (Vacuolar protein sorting-associated protein 11)	Q9H270	MULTISUBUNIT TETHERING COMPLEX	CORVET	27	941
Vps18/PEP3 (Vacuolar protein sorting-associated protein 18)	Q9P253	MULTISUBUNIT TETHERING COMPLEX	CORVET	45	973
EXOC5 (Exocyst complex component 5) (Sec10 hom)	O00471	MULTISUBUNIT TETHERING COMPLEX	Exocyst	13	708
EXOC3 (Exocyst complex component 3) (Sec6 hom)	O60645	MULTISUBUNIT TETHERING COMPLEX	Exocyst	38	756
EXOC8 (Exocyst complex component 8) (Exo84 hom)	Q8IYI6	MULTISUBUNIT TETHERING COMPLEX	Exocyst	131	725
EXOC6 (Exocyst complex component 6) (Sec15A)	Q8TAG9	MULTISUBUNIT TETHERING COMPLEX	Exocyst	36	804
EXOC4 (Exocyst complex component 4) (Sec8 hom)	Q96A65	MULTISUBUNIT TETHERING COMPLEX	Exocyst	62	974
EXOC2 (Exocyst complex component 2) (Sec5 hom)	Q96KP1	MULTISUBUNIT TETHERING COMPLEX	Exocyst	46	924
EXOC1 (Exocyst complex component 1) (Sec3 hom)	Q9NV70	MULTISUBUNIT TETHERING COMPLEX	Exocyst	102	894

EXOC7 (Exocyst complex component 7) (Exo70 hom)	Q9UPT5	MULTISUBUNIT TETHERING COMPLEX	Exocyst	90	735
EXOC6B (Exocyst complex component 1) (Sec15B)	Q9Y2D4	MULTISUBUNIT TETHERING COMPLEX	Exocyst	81	811
Vps53 (Vacuolar protein sorting-associated protein 53)	Q5VIR6	MULTISUBUNIT TETHERING COMPLEX	GARP	66	699
Vps52 (Vacuolar protein sorting-associated protein 52)	Q8N1B4	MULTISUBUNIT TETHERING COMPLEX	GARP	26	723
Vps54 (Vacuolar protein sorting-associated protein 54)	Q9P1Q0	MULTISUBUNIT TETHERING COMPLEX	GARP	87	977
Protein fat-free homolog (Ang2)	Q9UID3	MULTISUBUNIT TETHERING COMPLEX	GARP	87	782
Vps41 (Vacuolar protein sorting-associated protein 41)	P49754	MULTISUBUNIT TETHERING COMPLEX	HOPS	34	854
Vps39/VAM6 (Vacuolar protein sorting-associated protein 39)	Q96JC1	MULTISUBUNIT TETHERING COMPLEX	HOPS	22	886
ZW10 Centromere/kinetochore protein zw10 homolog	O43264	MULTISUBUNIT TETHERING COMPLEX	Tethering complexes	31	779
RAD50-interacting protein 1 (Tip20 homolog)	Q6NUQ1	MULTISUBUNIT TETHERING COMPLEX	Tethering complexes	24	792
TRAPPC3 Trafficking protein particle complex subunit 3 (Bet3 hom)	O43617	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	9	180
TRAPPC6A Trafficking protein particle complex subunit 6A	O75865	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	4	159
TRAPPC2 Trafficking protein particle complex subunit 2	P0DI81	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	0	140
TRAPPC2P1 Trafficking protein particle complex subunit 2 protein TRAPPC2P1	P0DI82	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	0	140
TRAPPC10 Trafficking protein particle complex subunit 10	P48553	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	81	1259
TRAPPC11 Trafficking protein particle complex subunit 11	Q7Z392	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	8	1133
TRAPPC6B Trafficking protein particle complex subunit 6B	Q86SZ2	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	0	158
TRAPPC5 Trafficking protein particle complex subunit 5	Q8IUR0	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	0	188
TRAPPC12 Trafficking protein particle complex subunit 12	Q8WVT3	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	275	735
TRAPPC9 Trafficking protein particle complex subunit 9	Q96Q05	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	69	1148
TRAPPC2L Trafficking protein particle complex subunit 2-like	Q9UL33	MULTISUBUNIT	TRAPPI	0	140

protein		TETHERING COMPLEX			
TRAPPC4 Trafficking protein particle complex subunit 4	Q9Y296	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	0	219
TRAPPC8 Trafficking protein particle complex subunit 8 (TRS85 hom)	Q9Y2L5	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	144	1435
TRAPPC1 Trafficking protein particle complex subunit 1 (Bet5 hom)	Q9Y5R8	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	0	145
CPLX1 Complexin-1	O14810	NEUROTRANSMISSION SPECIFIC REG.	Complexins (synaphins): SNARE binding proteins in Ca2+-regulated exocytosis	124	134
CPLX2 Complexin-2	Q6PUV4	NEUROTRANSMISSION SPECIFIC REG.	Complexins (synaphins): SNARE binding proteins in Ca2+-regulated exocytosis	132	134
CPLX4 Complexin-4	Q7Z7G2	NEUROTRANSMISSION SPECIFIC REG.	Complexins (synaphins): SNARE binding proteins in Ca2+-regulated exocytosis	122	160
CPLX3 Complexin-3	Q8WVH0	NEUROTRANSMISSION SPECIFIC REG.	Complexins (synaphins): SNARE binding proteins in Ca2+-regulated exocytosis	134	158
Synaptophysin (Major synaptic vesicle protein p38) (in regulated exocyt)	P08247	NEUROTRANSMISSION SPECIFIC REG.	Regulatory proteins	82	217
(Tomosyn, STXBP5) Syntaxin-binding protein 5 (in regulated exocyt)	Q5T5C0	NEUROTRANSMISSION SPECIFIC REG.	Regulatory proteins	211	1151
UNC13C (Protein unc-13 homolog C) (in regulated exocyt)	Q8NB66	NEUROTRANSMISSION SPECIFIC REG.	Regulatory proteins	603	2214
Munc18-1 (Syntaxin-binding protein 1) (SEC1 homolog)	P61764	NEUROTRANSMISSION SPECIFIC REG.	SM proteins	54	594

Synaptotagmin-5	O00445	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	63	365
Synaptotagmin-7	O43581	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	87	382
Synaptotagmin-1	P21579	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	61	399
Synaptotagmin-16	Q17RD7	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	315	645
Synaptotagmin-6	Q5T7P8	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	95	489
Synaptotagmin-10	Q6XYQ8	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	91	502
Synaptotagmin-13	Q7L8C5	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	72	403
Synaptotagmin-9	Q86SS6	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding,	119	470

			sensing proteins in Ca2+-regulated exocytosis)		
Synaptotagmin-12	Q8IV01	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	33	400
Synaptotagmin-2	Q8N9I0	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	79	398
Synaptotagmin-14	Q8NB59	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	141	532
Synaptotagmin-8	Q8NBV8	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	65	380
Synaptotagmin-3	Q9BQG1	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	215	569
Synaptotagmin-15	Q9BQS2	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	25	396
Synaptotagmin-17	Q9BSW7	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-	80	474

			regulated exocytosis)		
Synaptotagmin-11	Q9BT88	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	65	410
Synaptotagmin-4	Q9H2B2	NEUROTRANSMISSION SPECIFIC REG.	Synaptotagmins: (Ca2+ binding, sensing proteins in Ca2+-regulated exocytosis)	63	404
GOLGA2 (Golgin subfamily A member 2)	Q08379	OTHER FUSION REGULATOR	COPI SYSTEM	646	1002
CUX1 (Protein CASP)	Q13948	OTHER FUSION REGULATOR	COPI SYSTEM	172	657
GOLGB1 (Golgin subfamily B member 1)	Q14789	OTHER FUSION REGULATOR	COPI SYSTEM	1353	3238
GOLGA5 (Golgin subfamily A member 5/Golgin-84)	Q8TBA6	OTHER FUSION REGULATOR	COPI SYSTEM	454	710
USO1 (General vesicular transport factor p115)	O60763	OTHER FUSION REGULATOR	COPII SYSTEM	205	962
NSF (N-ethylmaleimide-sensitive fusion protein)	P46459	OTHER FUSION REGULATOR	Dissociation of Cis-SNARE complexes:	16	744
NAPA/SNAPA (Alpha-soluble NSF attachment protein)	P54920	OTHER FUSION REGULATOR	Dissociation of Cis-SNARE complexes:	3	295
NAPG/SNAPG (Gamma-soluble NSF attachment protein)	Q99747	OTHER FUSION REGULATOR	Dissociation of Cis-SNARE complexes:	30	312
NAPB/SNAPB (Beta-soluble NSF attachment protein)	Q9H115	OTHER FUSION REGULATOR	Dissociation of Cis-SNARE complexes:	18	298
UNC13B (Protein unc-13 homolog B) (in regulated exocyt)	O14795	OTHER FUSION REGULATOR	Regulatory proteins	306	1591
SNAPIN (SNARE-associated protein Snapin) (in regulated exocyt)	O95295	OTHER FUSION REGULATOR	Regulatory proteins	29	136
GATE-16 (Gamma-aminobutyric acid receptor-associated protein-like 2)	P60520	OTHER FUSION REGULATOR	Regulatory proteins	0	117
UNC13D (Protein unc-13	Q70J99	OTHER FUSION	Regulatory	143	1090

homolog D) (in regulated exocyt)		REGULATOR	proteins		
(Amisyn, STXBP6) Syntaxin-binding protein 6	Q8NFX7	OTHER FUSION REGULATOR	Regulatory proteins	27	210
UNC13A (Protein unc-13 homolog A) (in regulated exocyt)	Q9UPW8	OTHER FUSION REGULATOR	Regulatory proteins	327	1703
Munc18-2 (Syntaxin-binding protein 2)(SEC1 homolog)	Q15833	OTHER FUSION REGULATOR	SM proteins	35	593
Sec1 family domain-containing protein 1 (SLY1 homolog)	Q8WVM8	OTHER FUSION REGULATOR	SM proteins	39	642
Vps33b (in CORVET and HOPS complexes!)	Q9H267	OTHER FUSION REGULATOR	SM proteins	4	617
Vacuolar protein sorting-associated protein 45 (Vps45)	Q9NRW7	OTHER FUSION REGULATOR	SM proteins	21	570
SNAP23 (Synaptosomal-associated protein 23)	O00161	SNARE	Other t-SNARES	145	211
GOSR2/GC27 (Golgi SNAP receptor complex member 2/membrin)	O14653	SNARE	Other t-SNARES	71	191
BET1 (BET1 homolog)	O15155	SNARE	Other t-SNARES	4	97
GOSR1/GS28 (Golgi SNAP receptor complex member 1)	O95249	SNARE	Other t-SNARES	31	229
SNAP25 (Synaptosomal-associated protein 25)	P60880	SNARE	Other t-SNARES	128	206
BNIP1/SEC20 (Vesicle transport protein SEC20)	Q12981	SNARE	Other t-SNARES	20	207
VTI1A (Vesicle transport through interaction with t-SNAREs homolog 1A)	Q96AJ9	SNARE	Other t-SNARES	95	196
BET1L/GS15 (BET1-like protein)	Q9NYM9	SNARE	Other t-SNARES	1	90
USE1 (Vesicle transport protein USE1)	Q9NZ43	SNARE	Other t-SNARES	59	238
VTI1B (Vesicle transport through interaction with t-SNAREs homolog 1B)	Q9UEU0	SNARE	Other t-SNARES	109	211
YKT6 (Synaptobrevin homolog YKT6)	O15498	SNARE	Synaptobrevins (v-SNARE family)	13	198
VAMP4 (Vesicle-associated membrane protein 4)	O75379	SNARE	Synaptobrevins (v-SNARE family)	44	120
SEC22B (Vesicle-trafficking protein SEC22b)	O75396	SNARE	Synaptobrevins (v-SNARE	7	194

			family)		
VAMP5 (Vesicle-associated membrane protein 5)	O95183	SNARE	Synaptobrevins (v-SNARE family)	14	95
VAMP1 (Vesicle-associated membrane protein 1)	P23763	SNARE	Synaptobrevins (v-SNARE family)	37	98
VAMP7 (Vesicle-associated membrane protein 7)	P51809	SNARE	Synaptobrevins (v-SNARE family)	0	199
VAMP2 (Vesicle-associated membrane protein 2)	P63027	SNARE	Synaptobrevins (v-SNARE family)	34	96
VAMP3 (Vesicle-associated membrane protein 3/cellubrevin)	Q15836	SNARE	Synaptobrevins (v-SNARE family)	14	79
SEC22A (Vesicle-trafficking protein SEC22a)	Q96IW7	SNARE	Synaptobrevins (v-SNARE family)	0	226
SEC22C (Vesicle-trafficking protein SEC22c)	Q9BRL7	SNARE	Synaptobrevins (v-SNARE family)	0	219
VAMP8 (Vesicle-associated membrane protein 8)	Q9BV40	SNARE	Synaptobrevins (v-SNARE family)	13	79
STX16 (Syntaxin-16)	O14662	SNARE	Syntaxins (t-SNARE family)	51	304
STX7 (Syntaxin-7)	O15400	SNARE	Syntaxins (t-SNARE family)	105	240
STX6 (Syntaxin-6)	O43752	SNARE	Syntaxins (t-SNARE family)	41	234
STX10 (Syntaxin-10)	O60499	SNARE	Syntaxins (t-SNARE family)	18	228
STX11 (Syntaxin-11)	O75558	SNARE	Syntaxins (t-SNARE family)	21	287
STX2 (Syntaxin-2)	P32856	SNARE	Syntaxins (t-SNARE family)	66	264
STX17 (Syntaxin-17)	P56962	SNARE	Syntaxins (t-SNARE family)	30	281
STX1B (Syntaxin-1B)	P61266	SNARE	Syntaxins (t-SNARE	93	264

			family)		
STX4 (Syntaxin-4)	Q12846	SNARE	Syntaxins (t-SNARE family)	88	276
STX5 (Syntaxin-5)	Q13190	SNARE	Syntaxins (t-SNARE family)	72	334
STX3 (Syntaxin-3)	Q13277	SNARE	Syntaxins (t-SNARE family)	68	268
STX1A (Syntaxin-1A)	Q16623	SNARE	Syntaxins (t-SNARE family)	99	267
STX12 (Syntaxin-12)	Q86Y82	SNARE	Syntaxins (t-SNARE family)	121	255
STX19 (Syntaxin-19)	Q8N4C7	SNARE	Syntaxins (t-SNARE family)	18	294
STX18 (Syntaxin-18)	Q9P2W9	SNARE	Syntaxins (t-SNARE family)	114	314
STX8 (Syntaxin-8)	Q9UNK0	SNARE	Syntaxins (t-SNARE family)	40	219
Amphiphysin	P49418	UNCLASSIFIED	CLATHRIN	421	695
Endophilin-A2, EA2	Q99961	UNCLASSIFIED	CLATHRIN	131	368
Endophilin-A1	Q99962	UNCLASSIFIED	CLATHRIN	117	352
Endophilin-A3	Q99963	UNCLASSIFIED	CLATHRIN	33	347
SCYL1 (N-terminal kinase-like protein)	Q96KG9	UNCLASSIFIED	COPI SYSTEM	275	808
SH3GLB2 (Endophilin-B2)	Q9NR46	UNCLASSIFIED	COPI SYSTEM	40	395
SH3GLB1 (Endophilin-B1)	Q9Y371	UNCLASSIFIED	COPI SYSTEM	26	365
TMED3 (Transmembrane emp24 domain-containing protein 3)	Q9Y3Q3	UNCLASSIFIED	COPI SYSTEM	2	196
TMED10 (Transmembrane emp24 domain-containing protein 10)	P49755	UNCLASSIFIED	COPI SYSTEM/COPII SYSTEM	5	198
TMED2 (Transmembrane emp24 domain-containing protein 2)	Q15363	UNCLASSIFIED	COPI SYSTEM/COPII SYSTEM	12	180
TMED9 (Transmembrane emp24	Q9BVK6	UNCLASSIFIED	COPI SYSTEM/CO	3	215

domain-containing protein 9)			PII SYSTEM		
TMED7 (Transmembrane emp24 domain-containing protein 7)	Q9Y3B3	UNCLASSIFIED	COPI SYSTEM/COPII SYSTEM	2	203
Protein transport protein Sec16A	O15027	UNCLASSIFIED	COPII SYSTEM	1556	2179
YIF1A (Protein YIF1A)	O95070	UNCLASSIFIED	COPII SYSTEM	22	187
LMAN1 (Protein ERGIC-53)	P49257	UNCLASSIFIED	COPII SYSTEM	117	489
BCAP31 (B-cell receptor-associated protein 31)	P51572	UNCLASSIFIED	COPII SYSTEM	52	183
TMED1 (Transmembrane emp24 domain-containing protein 1)	Q13445	UNCLASSIFIED	COPII SYSTEM	2	206
VMA21 (Vacuolar ATPase assembly integral membrane protein VMA21)	Q3ZAQ7	UNCLASSIFIED	COPII SYSTEM	4	59
TMED4 (Transmembrane emp24 domain-containing protein 4)	Q7Z7H5	UNCLASSIFIED	COPII SYSTEM	2	209
MCFD2 (Multiple coagulation factor deficiency protein 2)	Q8NI22	UNCLASSIFIED	COPII SYSTEM	53	146
YIPF5 (Protein YIPF5)	Q969M3	UNCLASSIFIED	COPII SYSTEM	28	152
ERGIC1 (Endoplasmic reticulum-Golgi intermediate compartment protein 1), ERGIC32	Q969X5	UNCLASSIFIED	COPII SYSTEM	9	248
Protein transport protein Sec16B	Q96JE7	UNCLASSIFIED	COPII SYSTEM	576	1060
ERGIC2 (Endoplasmic reticulum-Golgi intermediate compartment protein 2)	Q96RQ1	UNCLASSIFIED	COPII SYSTEM	14	335
BCAP29 (B-cell receptor-associated protein 29)	Q9UHQ4	UNCLASSIFIED	COPII SYSTEM	56	178
ERGIC3 (Endoplasmic reticulum-Golgi intermediate compartment protein 3)	Q9Y282	UNCLASSIFIED	COPII SYSTEM	5	341
TMED5 (Transmembrane emp24 domain-containing protein 5)	Q9Y3A6	UNCLASSIFIED	COPII SYSTEM	0	208
SEC23IP/p125 (SEC23-interacting protein)	Q9Y6Y8	UNCLASSIFIED	COPII SYSTEM	282	1000

Table 2B. Proteins involved in the main trafficking pathways in yeast.

Protein name	Uniprot accession	Functional classification	System	Disordered residues	Total number of residues
PBI2 (Protease B inhibitors 2 and 1) (LMA-1 complex subunit)	P01095	OTHER FUSION REGULATORS	Regulatory proteins	15	75
YPT1 (GTP-binding protein YPT1)	P01123	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	22	206
ARF1 (ADP-ribosylation factor 1)	P11076	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	0	181
Vps11/PEP5 (Vacuolar protein sorting-associated protein 11)	P12868	MULTISUBUNIT TETHERING COMPLEX	CORVET	21	1029
SEC23 (Protein transport protein SEC23)	P15303	ADAPTOR/SORTING	COPII SYSTEM	35	768
Clathrin light chain	P17891	COAT	CLATHRIN	153	233
Arrestin-related trafficking adapter 10 (ART10)	P18634	ADAPTOR/SORTING	CLATHRIN	25	518
Sec18 (Vesicular-fusion protein SEC18, NSF homolog)	P18759	OTHER FUSION REGULATORS	Dissociation of Cis-SNARE complexes	33	758
Exo70	P19658	MULTISUBUNIT TETHERING COMPLEX	Exocyst	18	623
SAR1 (Small COPII coat GTPase SAR1)	P20606	ENZYME/ENZYME-INTERACTOR	COPII SYSTEM	0	190
Vps33 (in CORVET and HOPS complexes!)	P20795	MULTISUBUNIT TETHERING COMPLEX	SM proteins	3	691
Clathrin heavy chain	P22137	COAT	CLATHRIN	25	1653
Protein SLY1 (Sly1)	P22213	OTHER FUSION REGULATORS	SM proteins	66	666
SEC22 (Protein transport protein SEC22)	P22214	SNARE	Synaptobrevins (v-SNARE family)	0	193
TRX1 (Thioredoxin-1) (LMA-1 complex subunit)	P22217	OTHER FUSION REGULATORS	Regulatory proteins	0	103
Sec15	P22224	MULTISUBUNIT TETHERING COMPLEX	Exocyst	60	910
TRX2 (Thioredoxin-2) (LMA-1 complex subunit)	P22803	OTHER FUSION REGULATORS	Regulatory proteins	0	104
BET1 (Protein transport protein BET1)	P22804	SNARE	Other t-SNARES	28	118
Vps3 (Vacuolar protein sorting-associated protein 3)	P23643	MULTISUBUNIT TETHERING COMPLEX	CORVET	127	1011
BOS1 (Protein transport protein BOS1)	P25385	SNARE	Other t-SNARES	79	226
USO1 (Intracellular protein transport protein USO1)	P25386	OTHER FUSION REGULATORS	COPII SYSTEM	734	1790
Suppressor of yeast profilin deletion (SYP1)	P25623	ADAPTOR/SORTING	CLATHRIN	368	870
AP-2 complex subunit beta	P27351	ADAPTOR/SORTING	CLATHRIN	81	700
Vps18/PEP3 (Vacuolar protein sorting-associated protein 18)	P27801	MULTISUBUNIT TETHERING COMPLEX	CORVET	0	918
SEC20 (Protein transport protein SEC20)	P28791	SNARE	Other t-SNARES	19	366
Protein transport protein SEC1 (Sec1)	P30619	OTHER FUSION REGULATORS	SM proteins	114	724
SNC1 (Synaptobrevin homolog 1)	P31109	SNARE	Synaptobrevins (v-SNARE family)	26	100
SYN8 (Syntaxin-8)	P31377	SNARE	Syntaxins (t-SNARE family)	76	234
SEC21 (Coatomer subunit gamma)	P32074	ADAPTOR/SORTING	COPI SYSTEM	42	935
Actin cytoskeleton-regulatory complex protein PAN1	P32521	ADAPTOR/SORTING	CLATHRIN	1097	1480

Sec17 (Alpha-soluble NSF attachment protein)	P32602	OTHER FUSION REGULATORS	Dissociation of Cis-SNARE complexes	3	292
TCA17 (TRAPP-associated protein TCA17)	P32613	MULTISUBUNIT TETHERING COMPLEX	TRAPPII	0	152
Actin cytoskeleton-regulatory complex protein SLA1	P32790	ADAPTOR/SORTING	CLATHRIN	812	1244
EMP24 (Endosomal protein EMP24B)	P32803	UNCLASSIFIED	COPII SYSTEM	3	182
Sec6	P32844	MULTISUBUNIT TETHERING COMPLEX	Exocyst	6	805
PEP12 (Syntaxin PEP12)	P32854	SNARE	Syntaxins (t-SNARE family)	35	268
Sec8	P32855	MULTISUBUNIT TETHERING COMPLEX	Exocyst	101	1065
SSO1 (Protein SSO1)	P32867	SNARE	Syntaxins (t-SNARE family)	65	268
Trs65 (Trafficking protein particle complex II-specific subunit 65)	P32893	MULTISUBUNIT TETHERING COMPLEX	TRAPPII	57	560
VAM7 (Vacuolar morphogenesis protein 7)	P32912	SNARE	Other t-SNARES	111	316
SNC2 (Synaptobrevin homolog 2)	P33328	SNARE	Synaptobrevins (v-SNARE family)	27	96
Sec3 (Exocyst complex component SEC3)	P33332	MULTISUBUNIT TETHERING COMPLEX	Exocyst	465	1336
Protein SLA2, Transmembrane protein MOP2, END4; UFG1;	P33338	UNCLASSIFIED	CLATHRIN	256	948
Tip20 (Protein transport protein TIP20)	P33891	MULTISUBUNIT TETHERING COMPLEX	Tethering complexes	13	701
EDE1 (EH domain-containing and endocytosis protein 1)	P34216	ADAPTOR/SORTING	CLATHRIN	957	1381
AP-1 complex subunit sigma-1	P35181	ADAPTOR/SORTING	CLATHRIN	0	156
GCS1 (ADP-ribosylation factor GTPase-activating protein GCS1)	P35197	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	193	352
AP-1 complex subunit beta-1	P36000	ADAPTOR/SORTING	CLATHRIN	38	726
YKT6 (Synaptobrevin homolog YKT6)	P36015	SNARE	Synaptobrevins (v-SNARE family)	0	200
Vps51 (Ang2) (Vacuolar protein sorting-associated protein 51)	P36116	MULTISUBUNIT TETHERING COMPLEX	GARP	88	164
Arrestin-related trafficking adapter 6 (ALY1)	P36117	ADAPTOR/SORTING	CLATHRIN	312	915
Bet3 (Trafficking protein particle complex subunit BET3)	P36149	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	0	193
AP-2 complex subunit alpha	P38065	ADAPTOR/SORTING	CLATHRIN	48	1025
AP-3 complex subunit mu	P38153	ADAPTOR/SORTING	CLATHRIN	12	483
SRO77 (Lethal(2) giant larvae protein homolog SRO77, tomosyn hom)	P38163	OTHER FUSION REGULATORS	Regulatory proteins	75	1010
Autophagy-related protein 8 (Atg8, Apg8)	P38182	OTHER FUSION REGULATORS	Regulatory proteins	0	117
Exo84	P38261	MULTISUBUNIT TETHERING COMPLEX	Exocyst	244	753
Trs20 (Trafficking protein particle complex subunit 20)	P38334	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	19	175
GLO3 (ADP-ribosylation factor GTPase-activating protein GLO3)	P38682	ENZYME/ENZYME-INTERACTOR	COPI SYSTEM	297	493
Adaptin medium chain homolog APM2	P38700	ADAPTOR/SORTING	CLATHRIN	122	605

GOS1 (Golgi SNAP receptor complex member 1)	P38736	SNARE	Other t-SNARES	35	205
SFB3 (SED5-binding protein 3/SEC24-related protein 3/Lst1)	P38810	ADAPTOR/SORTING	COPII SYSTEM	135	929
ADP-ribosylation factor-binding protein, GGA2	P38817	ADAPTOR/SORTING	CLATHRIN	198	585
ERP5 (Protein ERP5)	P38819	UNCLASSIFIED	COPII SYSTEM	6	191
Clathrin coat assembly protein AP180A (YAP1801)	P38856	ADAPTOR/SORTING	CLATHRIN	307	637
SVP26 (Protein SVP26)	P38869	ADAPTOR/SORTING	COPII SYSTEM	6	149
Vacuolar protein sorting-associated protein 45 (Vps45)	P38932	OTHER FUSION REGULATORS	SM proteins	12	577
Vps41 (Vacuolar protein sorting-associated protein 41)	P38959	MULTISUBUNIT TETHERING COMPLEX	HOPS	134	992
SEC31 (Protein transport protein SEC31)	P38968	COAT	COPII SYSTEM	563	1273
AP-1 accessory protein LAA1	P39526	UNCLASSIFIED	CLATHRIN	19	2014
MST28 (Multicopy suppressor of SEC21 protein 28)	P39552	UNCLASSIFIED	COPI SYSTEM	24	192
Vps8 (Vacuolar protein sorting-associated protein 8)	P39702	MULTISUBUNIT TETHERING COMPLEX	CORVET	39	1274
ERP2 (Protein ERP2)	P39704	UNCLASSIFIED	COPII SYSTEM	1	194
ERV46 (ER-derived vesicles protein ERV46)	P39727	UNCLASSIFIED	COPII SYSTEM	29	373
Vps52 (Vacuolar protein sorting-associated protein 52)	P39904	MULTISUBUNIT TETHERING COMPLEX	GARP	48	641
SSO2 (Protein SSO2)	P39926	SNARE	Syntaxins (t-SNARE family)	103	273
Ddi1/Vsm1 (DNA damage-inducible protein 1/v-SNARE-master 1)	P40087	OTHER FUSION REGULATORS	Regulatory proteins	100	428
Cog3 (Sec34) (Conserved oligomeric Golgi complex subunit 3)	P40094	MULTISUBUNIT TETHERING COMPLEX	Cog	46	801
SEC9 (Protein transport protein SEC9)	P40357	SNARE	Other t-SNARES	606	651
SEC24 (Protein transport protein SEC24)	P40482	ADAPTOR/SORTING	COPII SYSTEM	176	926
SEC28 (Coatomer subunit epsilon)	P40509	COAT	COPI SYSTEM	0	296
TED1 (Protein TED1)	P40533	UNCLASSIFIED	COPII SYSTEM	5	431
Phosphatidylinositol 4,5-bisphosphate 5-phosphatase INP51 INP51	P40559	ENZYME/ENZYME-INTERACTOR	CLATHRIN	35	946
VMA21 (Vacuolar ATPase assembly integral membrane protein VMA21)	P41806	UNCLASSIFIED	COPII SYSTEM	9	34
SEC26 (Coatomer subunit beta)	P41810	ADAPTOR/SORTING	COPI SYSTEM	39	973
SEC27 (Coatomer subunit beta')	P41811	COAT	COPI SYSTEM	90	889
UFE1 (Syntaxin UFE1)	P41834	SNARE	Syntaxins (t-SNARE family)	28	328
EMP47 (Protein EMP47)	P43555	UNCLASSIFIED	COPII SYSTEM	43	424
RET2 (Coatomer subunit delta)	P43621	ADAPTOR/SORTING	COPI SYSTEM	158	546
SFT1 (Protein transport protein SFT1)	P43682	SNARE	Other t-SNARES	11	77

AP-3 complex subunit beta	P46682	ADAPTOR/SORTING	CLATHRIN	101	809
Trs85 (Trafficking protein particle complex III-specific subunit 85)	P46944	MULTISUBUNIT TETHERING COMPLEX	TRAPPII	59	698
Cargo-transport protein YPP1	P46951	UNCLASSIFIED	CLATHRIN	18	817
Arrestin-related trafficking adapter 3 (ALY2)	P47029	ADAPTOR/SORTING	CLATHRIN	549	1046
Vps53 (Vacuolar protein sorting-associated protein 53)	P47061	MULTISUBUNIT TETHERING COMPLEX	GARP	29	822
AP-3 complex subunit sigma	P47064	ADAPTOR/SORTING	CLATHRIN	1	194
ENT3 (Epsin-3)	P47160	ADAPTOR/SORTING	CLATHRIN	242	408
SEC16 (COPII coat assembly protein SEC16)	P48415	UNCLASSIFIED	COPII SYSTEM	1634	2195
Polyphosphatidylinositol phosphatase INP52 (Synaptojanin-like prot. 2)	P50942	ENZYME/ENZYME-INTERACTOR	CLATHRIN	283	1183
YIP1 (Protein transport protein YIP1)	P53039	OTHER FUSION REGULATORS	COPII SYSTEM	13	143
Cog1 (Conserved oligomeric Golgi complex subunit 1)	P53079	MULTISUBUNIT TETHERING COMPLEX	Cog	12	417
USE1/SLT1 (Protein transport protein USE1)	P53146	SNARE	Other t-SNARES	56	224
ERV14 (ER-derived vesicles protein ERV14)	P53173	UNCLASSIFIED	COPII SYSTEM	0	75
MST27 (Multicopy suppressor of SEC21 protein 27)	P53176	UNCLASSIFIED	COPI SYSTEM	17	192
Cog7 (Conserved oligomeric Golgi complex subunit 7)	P53195	MULTISUBUNIT TETHERING COMPLEX	Cog	38	279
ERP6, Protein ERP6	P53198	UNCLASSIFIED	COPII SYSTEM	0	195
Arrestin-related trafficking adapter 5 (ART5)	P53244	ADAPTOR/SORTING	CLATHRIN	100	586
VOA1 (V0 assembly protein 1)	P53262	UNCLASSIFIED	COPII SYSTEM	16	244
Cog2 (Sec35) (Conserved oligomeric Golgi complex subunit 2)	P53271	MULTISUBUNIT TETHERING COMPLEX	Cog	5	262
Clathrin coat assembly protein AP180B (YAP1802)	P53309	ADAPTOR/SORTING	CLATHRIN	293	568
ERV29 (ER-derived vesicles protein ERV29)	P53337	UNCLASSIFIED	COPII SYSTEM	2	184
RET3 (Coatomer subunit zeta)	P53600	ADAPTOR/SORTING	COPI SYSTEM	9	189
COP1/SEC33 (Coatomer subunit alpha)	P53622	COAT	COPI SYSTEM	91	1201
YIF1 (Protein transport protein YIF1)	P53845	OTHER FUSION REGULATORS	COPII SYSTEM	57	210
Dsl1 (Protein transport protein DSL1)	P53847	MULTISUBUNIT TETHERING COMPLEX	Tethering complexes	93	754
Cog5 (Conserved oligomeric Golgi complex subunit 5)	P53951	MULTISUBUNIT TETHERING COMPLEX	Cog	2	403
SFB2 (SED5-binding protein 2/SEC24-related protein 2/ ISS1)	P53953	ADAPTOR/SORTING	COPII SYSTEM	52	876
Cog6 (Conserved oligomeric Golgi complex subunit 6)	P53959	MULTISUBUNIT TETHERING COMPLEX	Cog	126	839
ERV25 (Endoplasmic reticulum vesicle protein 25)	P54837	UNCLASSIFIED	COPI SYSTEM	2	190
Sec5	P89102	MULTISUBUNIT TETHERING COMPLEX	Exocyst	65	971
AP-2 complex subunit sigma	Q00381	ADAPTOR/SORTING	CLATHRIN	0	147
AP-1 complex subunit mu-1-I	Q00776	ADAPTOR/SORTING	CLATHRIN	45	475

SED5 (Integral membrane protein SED5)	Q01590	SNARE	Syntaxins (t-SNARE family)	144	319
Vps16 (Vacuolar protein sorting-associated protein 16)	Q03308	MULTISUBUNIT TETHERING COMPLEX	CORVET	1	798
TLG1 (T-SNARE affecting a late Golgi compartment protein 1)	Q03322	SNARE	Syntaxins (t-SNARE family)	117	203
Trs31 (Trafficking protein particle complex subunit 31)	Q03337	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	35	283
Bet5 (Trafficking protein particle complex subunit BET5)	Q03630	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	3	159
Trs130 (Trafficking protein particle complex II-specific subunit 130)	Q03660	MULTISUBUNIT TETHERING COMPLEX	TRAPPII	2	1102
ENT5 (Epsin-5)	Q03769	ADAPTOR/SORTING	CLATHRIN	198	411
Trs23 (Trafficking protein particle complex subunit 23)	Q03784	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	39	219
Trs120 (Trafficking protein particle complex II-specific subunit 120)	Q04183	MULTISUBUNIT TETHERING COMPLEX	TRAPPII	14	1289
VTI1 (t-SNARE VTI1)	Q04338	SNARE	Other t-SNARES	37	196
SPO20 (Sporulation-specific protein 20)	Q04359	SNARE	Other t-SNARES	125	397
GRH1 (GRASP65 homolog protein 1)	Q04410	UNCLASSIFIED	COPII SYSTEM	91	372
SEC13 (Protein transport protein SEC13)	Q04491	COAT	COPII SYSTEM	22	297
Cog8 (Conserved oligomeric Golgi complex subunit 8)	Q04632	MULTISUBUNIT TETHERING COMPLEX	Cog	189	607
ERV41 (ER-derived vesicles protein ERV41)	Q04651	UNCLASSIFIED	COPII SYSTEM	0	310
ERP1 (Protein ERP1)	Q05359	UNCLASSIFIED	COPII SYSTEM	0	198
ENT2 (Epsin-2)	Q05785	ADAPTOR/SORTING	CLATHRIN	450	613
Cog4 (Conserved oligomeric Golgi complex subunit 4)	Q06096	MULTISUBUNIT TETHERING COMPLEX	Cog	38	861
Sec10	Q06245	MULTISUBUNIT TETHERING COMPLEX	Exocyst	31	871
ADP-ribosylation factor-binding protein GGA1	Q06336	ADAPTOR/SORTING	CLATHRIN	164	557
Auxilin-like clathrin uncoating factor SWA2	Q06677	UNCLASSIFIED	CLATHRIN	298	668
Vps39/VAM6 (Vacuolar morphogenesis protein 6)	Q07468	MULTISUBUNIT TETHERING COMPLEX	HOPS	11	1049
TLG2 (T-SNARE affecting a late Golgi compartment protein 2)	Q08144	SNARE	Syntaxins (t-SNARE family)	76	376
AP-3 complex subunit delta	Q08951	ADAPTOR/SORTING	CLATHRIN	223	932
AP-1 complex subunit gamma-1	Q12028	ADAPTOR/SORTING	CLATHRIN	20	832
SRO7 (Lethal(2) giant larvae protein homolog SRO7, tomosyn hom)	Q12038	OTHER FUSION REGULATORS	Regulatory proteins	86	1033
Vps54 (Vacuolar protein sorting-associated protein 54)	Q12071	MULTISUBUNIT TETHERING COMPLEX	GARP	135	889
BUG1 (Binder of USO1 and GRH1 protein 1)	Q12191	UNCLASSIFIED	COPII SYSTEM	288	341
RUD3 (GRIP domain-containing protein RUD3), GRP1	Q12234	UNCLASSIFIED	COPII SYSTEM	289	484
VAM3 (Syntaxin VAM3)	Q12241	SNARE	Syntaxins (t-SNARE family)	121	262

NYV1 (Vacuolar v-SNARE NYV1)	Q12255	SNARE	Synaptobrevins (v-SNARE family)	21	232
Polyphosphatidylinositol phosphatase INP53 (Synaptojanin-like prot. 3)	Q12271	ENZYME/ENZYME-INTERACTOR	CLATHRIN	235	1107
EMP46 (Protein EMP46)	Q12396	UNCLASSIFIED	COPI SYSTEM	19	423
ERP3 (Protein ERP3)	Q12403	UNCLASSIFIED	COPII SYSTEM	23	204
ERP4 (Protein ERP4)	Q12450	UNCLASSIFIED	COPII SYSTEM	0	186
ENT1 (Epsin-1)	Q12518	ADAPTOR/SORTING	CLATHRIN	303	454
Dsl3(Sec39) (Protein transport protein SEC39)	Q12745	MULTISUBUNIT TETHERING COMPLEX	Tethering complexes	10	709
YOS1 (Protein transport protein YOS1)	Q3E834	OTHER FUSION REGULATORS	COPII SYSTEM	1	43
AP-2 complex subunit mu	Q99186	ADAPTOR/SORTING	CLATHRIN	20	491
Trs33 (Trafficking protein particle complex subunit 33)	Q99394	MULTISUBUNIT TETHERING COMPLEX	TRAPPI	35	268